1-1-1979

Applications of latent trait theory to the development of norm-referenced tests.

Linda L. Cook
University of Massachusetts Amherst

APPLICATIONS OF LATENT TRAIT THEORY TO THE DEVELOPMENT OF NORM-REFERENCED TESTS

A Dissertation Presented

By

LINDA LEE COOK

Submitted to the Graduate School of the
University of Massachusetts in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

September 1979

EDUCATION

APPLICATIONS OF LATENT TRAIT THEORY TO THE

DEVELOPMENT OF NORM-REFERENCED TESTS

A Dissertation Presented

By

LINDA LEE COOK

Approved as to style and content by:

Ronald K. Hambleton

Dr. Ronald K. Hambleton, Chairperson

Hariharan Swaminathan

Dr. Hariharan Swaminathan, Member

Dr. Harry Schumer, Member

Dr. Mario Fantini, Dean
School of Education

TO:

Peter and Perry

A C K N O W L E D G E M E N T S

Without the assistance, encouragement, and support of many individuals this dissertation could never have been completed. I wish to convey my deep appreciation to my three committee members, Dr. Ronald Hambleton, Dr. Hariharan Swaminathan, and Dr. Harry Schumer. Special recognition must be extended to my chairman, Dr. Hambleton, for his intellectual guidance and emotional support; for the true friendship he so willingly extended; for his unselfish gift of his personal time; and most of all for his patience. Dr. Swaminathan deserves special thanks for his guidance, friendship, and encouragement during my early years as a graduate student. Dr. Schumer was the professor of the first subject I studied at the University of Massachusetts many years ago as an undergraduate and for this and other reasons assumes particular importance to me as a member of my dissertation committee. I am indebted to my graduate colleagues in the Laboratory of Psychometric and Evaluative Research who have been an unflagging source of encouragement and friendship, particularly to Ms. Janice Gifford who has been a valuable friend and counselor but also assisted greatly in the many computer simulations that were required by this study. I also wish to express my gratitude to Ms. Bernadette McDonald who, through her skillful typing and editing added immeasurably to the quality of this dissertation. I am deeply indebted to my parents,

Mrs. Jessie Dean and Mr. Howard Dean for their sustained support and caring. Finally, I would like to acknowledge the encouragement and understanding I have received from my good friend and fellow graduate student, Daniel Eignor who with a delicate combination of humor and criticism, always managed to help me restore perspective at times when I felt truly defeated.

ABSTRACT

Applications of Latent Trait Theory to the

Development of Norm-Referenced Tests

(September 1979)

Linda Lee Cook, B.S., Ursinus College
M.E.D., Smith College
Ed.D., University of Massachusets, Amherst

Directed by: Ronald K. Hambleton

Latent trait theory offers several advantages to the psychometrician interested in developing tests: (1) invariant item parameters that facilitate the test development process as well as make possible the development of tests for a variety of applications, (2) a mathematical function that can be manipulated to provide valuable insights into how examinees perform on specific test items, and (3) added information about examinee ability derived from new test scoring methods. Because of these and other properties of latent trait models, item selection and item analysis processes differ substantially from those employed when using standard testing tehcnology.

Although the above mentioned advantages have been documented in the literature, to date, no specific methodology has been set forth for the development of norm-referenced tests utilizing latent trait theory. In part this is because there still exist a number of significant problem areas requiring resolution. For example, before

latent trait theory can be used successfully in test development work, more needs to be known about: (1) the robustness of the models, (2) the properties of information functions, and (3) how best to use these functions in the test development process.

This study had three purposes. The first was to study, systematically, the "goodness of fit" of the one-, two-, and three-parameter logistic models employing a practical criterion for assessment. Using computer-simulated test data, the effects of four variables were studied: (1) variation in item discrimination parameters, (2) the average value of the psuedo-chance level parameters, (3) test length, and (4) the shape of the ability distribution.

The second purpose of the study was to address two practical questions which are of importance and interest to test developers:

1. What are the effects of examinee sample size and test length on the precision of the standard error of ability estimation curves?

2. What effects do the statistical characteristics of an item pool have on the precision of standard error of ability estimation curves?

As in the previous study, computer-simulated test data was used to study the problem.

The third purpose of the study was to investigate the following questions related to the development of item selection methodologies:

1. Using a typical item pool (where items are described by parameters in the three-parameter logistic test model), how does one develop alternate item selection methodologies and how do the score information curves that result from these methodologies compare?

2. Given a specific testing purpose such as producing a scholarship exam or a test to optimally separate examinees into three ability categories, how does one develop alternate item selection methodologies and how do the score information curves resulting from these methodologies compare?

The results of the robustness studies revealed that there are some sizeable gains to be expected with modest length tests (n=20) in the correct ordering of examinees at the lower end of the ability continuum when three-parameter model estimates are used (as opposed to number right score). The gains were cut roughly in half when the tests were doubled (n=40) in length. It was also noted that item discrimination parameters as weights had little effect on the results.

Results from the second part of the study indicated that: (1) both test length and sample size are extremely important factors in the precision of SEE curves; (2) the precision of SEE curves at the extremes of an ability continuum is very poor, even with large examinee sample sizes, however, the results are substantially better when tests are lengthened, even if sample size is small; (3) the precision of SEE curves would be acceptable in most instances if the curves are based on 200 or more examinees with tests with at least 20 items, and; (4) the most sizeable improvements in the precision of SEE curves occur when examinee sample size is increased from 50 to 200 and when test length is increased from 10 to 20 items.

The third part of the study revealed that in all cases, item selection methods based on either random selection of items or the use of classical item statistics produced results inferior to those produced by methods utilizing latent trait model item parameters. The study also indicated that methods must be developed with a specific testing purpose in mind. If maximum information is required at only one point on an ability continuum, it is clear that

a method that chooses items that maximize information at this particular point will be the best. If information is required over a wider range of abilities, methods involving averaging the information values across ability levels of interest or choosing items in some systematic way that considers each point of interest on the ability continuum appear to be quite promising.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION AND STATEMENT OF THE PROBLEM

1.1 Background and Review of the Literature

There are many well-documented shortcomings of standard testing and measurement technology. For one, the values of standard item parameters (item difficulty and item discrimination) are not invariant across groups of examinees that differ in ability. This means that standard item statistics are only useful in test construction for examinee populations very similar to the sample of examinees in which the item statistics were obtained.

Another shortcoming of standard testing technology is that estimates of an examinee's ability depend on the specific set of test items administered to that examinee. Therefore, comparisons of examinee ability are only meaningful in situations where examinees are administered the same test items, parallel test items, or items that have been carefully equated. The fact that tests that have been developed employing standard testing technology produce ability estimates that depend on a specific set of items presents a particular problem for those interested in tailored testing. Tailored tests are designed such that test items are administered to examinees that are carefully selected to "match" their ability levels (Lord, 1970, 1974a; Weiss, 1976; Wood, 1973). In "tailored testing," it is likely that no two examinees will take the same set

of test items (or even the same number of test items). Since some examinees will be administered more difficult sets of test items than other examinees, the usual examinee test scores do not provide an adequate basis for ranking examinees on the ability measured by the test items.

Besides the two shortcomings of standard testing technology mentioned above, standard testing technology has failed to provide satisfactory solutions to many testing problems (for example, test design, test score equating, and item bias). For these and other reasons, many psychometricians have been investigating and developing more appropriate theories of mental measurement. Consequently, considerable attention is being currently directed toward the field of *latent trait theory*.

Latent trait theory can be traced back to the work of Lawley (1943, 1944). Lazarsfeld (1950) was perhaps the first to introduce the term "latent traits." The work of Lord (1952, 1953a, 1953b), however, is generally regarded as the "birth" of latent trait theory (or modern test theory as it is sometimes called). Progress in the 1950's and 60's was painstakingly slow, in part due to the mathematical complexity of the field, the lack of convenient and efficient computer programs to analyze the data according to latent trait theory, and the general skepticism about the gains that might accrue from this particular line of research. However, important breakthroughs recently in problem areas such as test score equating (Lord, 1975a; Rentz & Bashaw, 1975), tailored testing (Lord, 1974a; Weiss, 1976), test and design and test evaluation (Wright, 1968) through

applications of latent trait theory, have attracted considerable interest from measurement specialists. Other factors that have contributed to the current interest in latent trait theory include the availability of a number of useful computer programs and publication of a variety of successful applications in measurement journals (Bock, 1972; Lord, 1968, 1974a, 1975c; Samejima, 1969, 1972; Whitely & Dawis, 1974; Wright & Panchapakesan, 1969).

A theory of latent traits supposes that in testing situations, examinee performance on a test can be predicted (or explained) by defining characteristics of examinees, referred to as *traits*, and using the scores to predict or explain test performance (Lord & Novick, 1968). Since the traits are not directly measurable and therefore "unobservable," they are often referred to as *latent traits* or *abilities*. A latent trait model specifies a relationship between observable examinee test performance and the unobservable traits or abilities assumed to underlie performance on the test. The relationship between the "observable" quantites is described by a *mathematical function*. For this reason, latent trait models are *mathematical models*. Latent trait models are based on a number of assumptions concerning the test data. When selecting a particular latent trait model to apply to one's test data, it is necessary to consider whether the test data satisfy the assumptions of the model. If they do not, different test models should be considered. Alternately, some psychometricians (for example, Wright, 1968) have recommended that test developers design their tests so as to satisfy the assumptions of the particular latent trait models they are interested in using.

If latent trait theory is to fulfill the potential it holds for the field of educational and psychological measurement, a method for developing tests by applying the theory must be established. As the state of the art exists, there is only one well-defined and field tested methodology for developing tests, i.e., the application of classical test theory methods to the development of tests. Many theoreticians have been advocating the use of latent trait theory in the development of these types of tests. However, no specific methodology that can be followed by the practitioner exists.

Latent trait theory offers several advantages to the psychometrician interested in developing tests: For example, (1) invariant item parameters that facilitate the test development process as well as make possible the development of tests for a variety of applications, (2) a mathematical function that can be manipulated to provide valuable insights into how examinees perform on specific test items, and (3) added information about examinee ability derived from new test scoring methods (Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1977). Because of these and other properties of latent trait models, item selection, and item analysis processes differ substantially from those employed when using standard testing technology.

The following is a brief description of some of the important ways in which latent trait theory can facilitate the test development process: (1) statistics used to describe test items will not depend on the ability distribution of the specific group used to calibrate them (Lord and Novick, 1968); (2) when latent trait item parameters are known, the psychometrician can examine the contribution

of each test item to the test information curve, thus enabling the test developer to build a test which precisely fulfills a set of desired test specifications; (3) a psychometrician can form different combinations of items (tentative tests) in the initial stages of test development and compare the information curves of different sets of items at specific ability levels thus allowing him/her to choose the set of items most suited for the intended purpose of the test (Marco, 1977); (4) latent trait theory provides a method of examining item bias (Pine, 1976; Wright, Mead & Draba, 1976) which unlike classical test theory methods for studying the problem, is not affected by the difference in the ability levels of examinee groups being investigated.

1.2 Statement of the Problems

In view of the many successful applications of latent trait theory to a variety of mental measurement problems, the issue of whether or not to use latent trait theory seems to be resolved. However, latent trait theory is still relatively new and hence there remain many problem areas that need to be researched so as to increase the chance of successful application of the theory to test development. Three of the problem areas most important to the test development process focus on (1) the robustness of latent trait models; (2) the stability of item information functions; and (3) the use of item and test information functions for item selection.

The results of several studies have been reported that relate to the question of model robustness (Dinero & Haertel, 1977; Hambleton, 1969; Hambleton & Traub, 1976; Panchapakesan, 1969; Cook & Eignor, 1979). The findings have been quite contradictory, perhaps in some instances because of the confounding effects of sample size.

The basic problem with most of the goodness-of-fit and robustness studies that have been conducted recently is that they do not provide the practitioner with information that he/she may use when applying latent trait theory to the test development process. It is important for practitioners to see comparisons of the fit of latent trait models to various data sets using a criterion measure that has some practical meaning to them. To date there have been no comparative studies of the various latent trait models using practical criteria to judge the results.

The increasing use of test information functions as a means of constructing and evaluating tests, has been documented in most of the current literature concerning educational and psychological measurement. However, some important questions remain to be answered before these functions can be optimally applied to produce the desired results. These questions address the stability of the functions under varying circumstances. Variables unique to each testing situation, such as the characteristics of the item pool, the number of examinees used to estimate the parameters of the items contained in the pool and the number of items comprising the test will be reflected in the accuracy of the estimate of test information. Therefore, it seems apparent that an investigation of the influence of

these variables on the stability of information functions would be useful to those interested in using these functions as part of the test development process.

Lord (1977) discussed a procedure, outlined by Birnbaum (1968), for building a test utilizing item information functions. This procedure consists of the following steps:

1. Decide on the purpose of the test. Based on this purpose, determine the standard error of estimate required at each ability level and consequently the target information curve.

2. Select items with item information curves that fill hard to fill areas under the target information curve.

3. Continue to select items until the test information curve approximates the target information curve with the desired degree of accuracy.

The major problem with this procedure concerns the fact that it is not sufficiently operationalized to be applied by the practitioner interested in using information functions to build tests. For example, a specific methodology must be developed for: (1) establishing target information curves that are suitable for various testing purposes, and (2) selecting items such that the fewest number of items will be selected in the most efficient manner. It seems apparent that the operationalization of Birnbaum's procedure, including the development of algorithms for item selection, would greatly expedite the application of latent trait theory to the test development process.

1.3 Purposes

The previous section of this chapter has delineated a number of areas that require research before latent trait theory can be successfully applied to the development of norm-referenced tests. The research presented in this dissertation concentrates on three specific areas.

The focus of the first area of research that is described in this thesis was the robustness of latent trait models. The purpose of this research was to study the "goodness-of-fit" of the one-, two-, and three-parameter models employing practical criteria for assessment.

The second part of the study investigated the stability of test information functions. The concerns of this study were to systematically investigate:

1. The effects of examinee sample size and test length on the precision of test information functions.
2. The effects of the statistical characteristics of an item pool on test information functions.

The purpose of this part of the study was to provide guidelines to aid test developers in determining the confidence they should have in the test information functions that they utilize in their work.

The third area of research addressed in this thesis was the operationalization of Birnbaum's procedure for the use of item information functions for item selection. The purpose of this research was to develop a set of guidelines to be used by the practitioner when making decisions concerning a number of practical problems that arise when employing information functions to build

tests. In order to accomplish this purpose, studies were carried out that focused on the development and comparison of item selection algorithms suited for specific test construction purposes. This study investigated how best to establish a target information curve and compared several algorithms for selecting items to fit information curves.

1.4 Organization of the Study

The next five chapters of this thesis are organized in the following manner. Chapter II provides the theoretical framework for the research. Chapter III presents the robustness studies. Chapter IV contains the studies concerning the stability of item information curves, and Chapter V contains the studies related to the operationalization of Birnbaum's procedure. Chapters III through V are self-contained and share the following format:

1. Introduction
2. Methods of Investigation
3. Results and Discussion
4. Conclusion

The sixth and final chapter in this thesis is devoted to a summary of the studies as well as conclusions and suggestions for further research.

CHAPTER II

LATENT TRAIT MODELS AND RELATED CONCEPTS

2.1 Introduction

The purpose of this chapter is to introduce the topic of latent trait models. First, a brief non-mathematical introduction to the theory of latent traits will be provided. Second, the features of three latent trait models that seem to be particularly appropriate for use with mental test data will be reviewed. Third, the classical test model will be compared with latent trait models.

2.2 Features of Latent Trait Models

There are at least three fundamental notions in the general theory of latent traits: The *dimensionality of the latent space*, *local independence*, and *item characteristic curves*. Each of these notions will be discussed briefly below.

The *dimensionality of the latent space* refers to the number of latent traits that underlie examinee test performance. It is typical to assume that the latent space is unidimensional; that is, assume that the items in a test are homogeneous in the sense of measuring only a single ability or latent trait. Latent trait models in which the unidimensional assumption is not made are complex, and to date, not well developed. According to Lord (1968), the assumption

concerning the unidimensional nature of a set of items is not strictly true for most tests. However, he adds that it may provide a tolerably good approximation in some instances. The appropriateness of the assumption of unidimensionality for any set of mental test data can be partially studied through a factor analysis of the test items. (For details of one attempt at this, the reader is referred to Hambleton and Traub [1973].) When the items in a test measure more than a single ability, the items can be clustered into homogeneous groups on the basis of the results from a factor-analytic study. Then, a latent trait analysis can be applied to each homogeneous cluster of items. All further discussions of latent trait models in this Chapter will be restricted to models that assume a single ability underlying test performance.

The second notion is the *principle of local independence*. There are two forms of this principle, referred to as the *strong* and *weak* form of the principle of local independence. The strong form of the principle states that the test item responses of each examinee are statistically independent. This means, for example, that the probability of any examinee response pattern across a set of test items is given by the product of probabilities representing success on *each* item for that examinee. Also, it means that examinee performance on one test item does not affect the examinee's success or failure on any other item in the test.

The weak form of the principle is obtained by substituting "uncorrelated" for "statistically independent" in the statement of the principle. The distinction between the two forms of the principle

is the same one we often make in correlational research: we distinguish between variables being statistically independent and variables being uncorrelated; the first condition being a stronger statement about the relationship between two variables than the second.

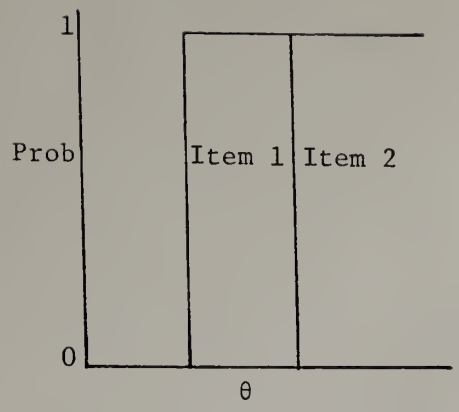
It is relatively easy to see that the assumption of local independence and the assumption of a unidimensional latent space are identical. To say that a single ability accounts for examinee performance on a set of test items is the same thing as saying that for examinees at the same ability level, their item responses are statistically independent, i.e., satisfy the principle of local independence. If this were not the case (i.e., if examinee item responses were statistically dependent), then it would follow that at least one more ability was being measured by the test items. The interested reader is referred to Lord and Novick (1968) for further clarification of this point.

The principle of local independence represents a restrictive assumption and so may not be satisfied with many sets of mental test data. Because of the equivalence of the principle of local independence and the assumption of unidimensionality, the appropriateness of the principle of local independence with any data set can also be tested, in part, using factor analytic techniques.

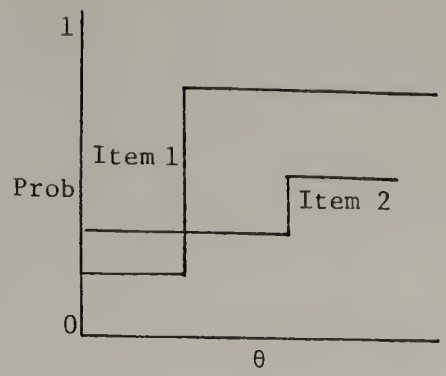
It should be recognized that the principle of local independence does *not* imply that test items are uncorrelated over the total group of examinees (Lord & Novick, 1968, p. 361). Positive correlations between pairs of items will result whenever there is variation among the examinees on the ability measured by the test items.

The third notion is that of an *item characteristic curve* (sometimes referred to as a *trace line*, or an *item characteristic function* when the latent space is multidimensional, i.e., when the number of latent traits underlying test performance exceeds one). An item characteristic curve is a mathematical function that relates the probability of success on an item to the ability measured by the test. A primary distinction among various latent trait models is in the mathematical form of the item characteristic curve. Examples of the mathematical forms of item characteristic curves of six latent trait models are shown in Figure 2.2.1. Each item characteristic curve for a particular latent trait model is a member of a family of curves of the same general form. For example, the item characteristic curve of the latent linear model (Figure 2.2.1, C) has the general form $P_g(\theta) = b_g + a_g\theta$, where $P_g(\theta)$ designates the probability of a correct response by an examinee with ability level θ, on an item g that is described by two parameters, denoted a_g and b_g. An item characteristic curve is specified completely when the general form is specified and the parameters of the curve for a particular item are known. The number of parameters required to describe an item characteristic curve will depend on the particular latent trait model. It is common though for the number of parameters to be one, two, or three.

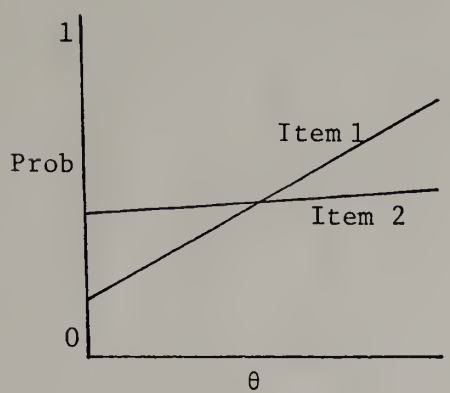
While item characteristic curves in the latent linear model are all straight lines (a restriction placed on us when we select the latent linear model), across different items in the test, the "curves" (or lines in this particular case) will vary in their intercepts and slopes to reflect the fact that the test items vary in "difficulty" and "discriminating power."



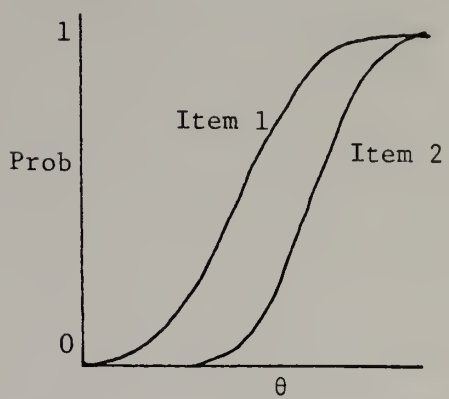
(a) perfect scale curves



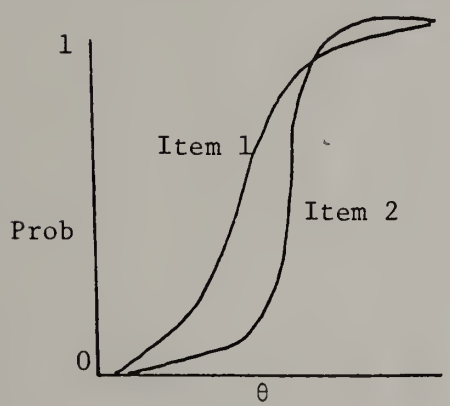
(b) latent distance curves



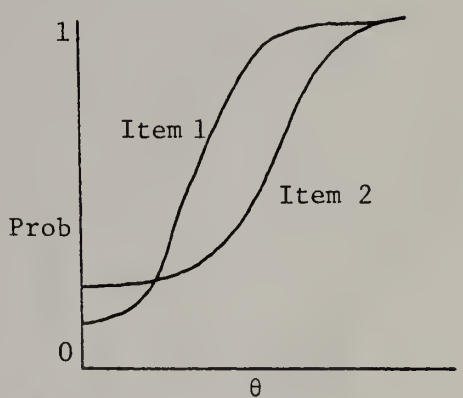
(c) latent linear curves



(d) one-parameter logistic curves



(e) two-parameter logistic curves



(f) three-parameter logistic cirves

Figure 2.2.1. Six examples of item characteristic curves.

In any practical application of latent trait models, it is usually necessary to specify the mathematical form of the item characteristic curves and obtain estimates of the item parameters needed to describe the curves. Readers are referred to Bock (1972), Lord (1968, 1974a), Whitely and Dawis (1974), Wright and Panchapakesan (1969), and Hambleton, Swaminathan, Cook, Eignor, and Gifford (1979) for details on some of the current methods for estimating item characteristic curve parameters (and ability estimates as well) of some of the more popular latent trait models.

An item characteristic curve represents the probability of a correct answer to an item expressed as a function of ability. However, the probability of a correct answer to an item is obviously *independent* of the distribution of examinee ability in the population of examinees of interest. Clearly, the probability of a correct response for an examinee will not depend on how many other examinees are located at the same location on the ability continuum. Therefore, the shape of an item characteristic curve does *not* depend on the distribution of ability in the examinee population. In some sense then, the shape of the curve will be invariant across different samples of examinees from that population, regardless of how the sample of examinees is selected. This important point will be expanded on later.

2.2.1 Common Forms of Item Characteristic Curves

In this section, three mathematical functions that are commonly used to represent item characteristic curves will be introduced. All three functions can be applied to binary-scored items administered

under non-speeded conditions. (When the items are administered under speeded conditions, it becomes necessary to distinguish between "omitted" items and "not reached" items by examinees so as to properly estimate examinee ability scores.)

(a) Two-Parameter Logistic Curves

Birnbaum (1968) proposed a latent trait model in which the item characteristic curve takes the form of a two-parameter logistic distribution function,

$$P_g(\theta) = \frac{e^{Da_g(\theta-b_g)}}{1+e^{Da_g(\theta-b_g)}} \quad (g = 1, 2, \ldots, n). \qquad [2.2.1.1]$$

In this equation, $P_g(\theta)$ is the probability that an examinee with ability θ answers item g correctly, a_g and b_g are parameters for item g(g=1,2,...,n) and n is the number of items in the test. The parameter b_g is usually referred to as the index of *item difficulty*. It represents the point on the ability scale at which the slope of the item characteristic curve is a maximum. The parameter, a_g, called item discrimination, is proportional to the slope of $P_g(\theta)$ at the point $\theta=b_g$. The constant D is a scaling factor. Usually we take D=1.7, to maximize the agreement between the logistic model and the normal-ogive model, a model that was originally studied by Lord (1952) but is mathematically inconvenient to work with.

The item difficulty parameter, b_g, is defined on the same scale as ability $[-\infty, +\infty]$. In practice though the range of b_g is from about -2 to +2 (assuming the ability distribution is centered with a mean equal to zero and standard deviation equal to one).

As b_g takes on values from -2 to +2, the items move from being very easy to very difficult for the group of examinees.

The item discrimination parameter, a_g, is defined, theoretically, on the scale $[-\infty, +\infty]$. However, negatively discriminating items are discarded from ability tests, and it therefore is unusual to obtain a_g values larger than two. High values of a_g result in item characteristic curves that are very "steep." Low values of a_g lead to item characteristic curves that increase gradually as a function of ability.

Careful inspection of the two-parameter logistic model reveals an additional implicit assumption characteristic of most latent trait models: guessing does not occur. That this must be so is apparent from the fact that as long as $a_g>0$ (that is, as long as there is a positive relationship between performance on the test item and the ability measured by the test), the probability of a correct response to an item decreases to zero as ability decreases.

(b) Three-Parameter Logistic Model

The three-parameter model is obtained from the two-parameter model by adding a third parameter, denoted c_g. The mathematical form of the three-parameter logistic curve is written

$$P_g(\theta) = c_g + (1-c_g)\frac{e^{Da_g(\theta-b_g)}}{1+e^{Da_g(\theta-b_g)}} \quad (g = 1, 2, \ldots, n). \quad [2.2.1.2]$$

The parameter c_g is the lower asymptote of the item characteristic curve and represents the probability of low ability examinees correctly answering a question. The purpose of including a paramter c_g

into the model is to attempt to account for the misfit of item characteristic curves at the low end of the ability continuum, where among other things, guessing is a factor in test performance. It has been common to refer to the parameter c_g as the *guessing* parameter in the model.

It is perhaps surprising to note then that typically the parameter c_g takes a value smaller than the value corresponding to the probability of a correct answer to a test item from random guessing. As Lord (1974b) has noted, this event is probably due to the ingenuity of item writers in developing "attractive" but incorrect choices. For this reason, discontinuation of the label "guessing parameter" to describe the parameter c_g would seem to be desirable.

(c) One-Parameter Logistic Model (Rasch Model)

Many researchers have become aware of the work of Georg Rasch, a Danish mathematician, in the area of latent trait models (Rasch, 1966), both through his own publications and the papers of others advancing his work (Anderson, Kearney, & Everett, 1968; Wright, 1968; Wright & Panchapakesan, 1969). Although the Rasch model was developed independently of other latent trait models and along quite different lines, Rasch's model can be viewed as a latent trait model in which the item characteristic curve is a one-parameter logistic function. Consequently, Rasch's model is a special case of Birnbaum's two-parameter logistic model, in which all items are assumed to have equal discriminating power and vary only in terms of difficulty. The form of the item characteristic curve for this model can then be written as

$$P_g(\theta) = \frac{e^{D\bar{a}(\theta-b_g)}}{1+e^{D\bar{a}(\theta-b_g)}} \quad (g = 1, 2, \ldots, n), \qquad [2.2.1.3]$$

in which $\bar{a}$, the only term not previously defined, is the common level of discrimination for all the items and is often set equal to one.

2.3 The Ability Scale and Its Meaning

That there is a more basic scale of ability than the true score scale for a test is obvious when it is recognized that the true score distributions (and observed score distributions) of non-parallel measures of a common ability will differ. The ability scale for a particular latent trait model is defined such that the distribution of abilities in a group of examinees will be identical regardless of the particular test measuring the ability (Lord, 1975a).

The ability scale is chosen so that the relationship between ability scores and item responses can be represented by item characteristic curves of some specified mathematical form. The ability scale is "stretched" and "compressed" at different points so as to maximize the "fit" between the item responses, item characteristic curves, and the ability scores. The resultant ability scale is unique up to the origin and unit of measurement which are arbitrary.

2.4 Test Information and Efficiency

Once a latent trait model is specified, the precision with which it estimates examinee ability can be determined. Of coures, the validity of the results will depend on the match between the model

and the test data. Following Sir Ronald Fisher's important statistical work in the 1920's, Birnbaum (1968) defined the notion of information as a quantity inversely proportional to the squared length of the confidence interval around an estimate of an examinee's ability. The standard error of estimate of ability is equal to $1/\sqrt{\text{information}}$. When information at an ability level is high, we have narrow confidence bands. Because the information function varies with ability level, it has been suggested that test information curves ought to replace the use of classical reliability estimates and standard errors of measurement in test score interpretations.

In mathematical terms, Birnbaum (1968) gives the information curve of a given scoring formula by

$$I_y(\theta) = \frac{\left[\sum_{g=1}^{n} w_g P'_g\right]^2}{\sum_{g=1}^{n} w^2_g P_g Q_g} \qquad [2.4.1]$$

In the expression above, $I_y(\theta)$ is the amount of information at ability level θ provided by the scoring formula y, where

$$Y = \sum_{g=1}^{n} w_g X_g \; ; \qquad [2.4.2]$$

the variable X_g is 0 or 1 depending on whether or not item g is answered correctly; P_g is the probability of a correct answer to item g by an examinee with ability level θ; Q_g is equal to $1 - P_g$; P'_g is the slope of the item characteristic curve at ability level θ; and the item scoring weights are w_g, g = 1, 2, ..., n.

Birnbaum (1968 has shown that the maximum value of $I_y(\theta)$, referred to as the test information curve, is given by

$$I(\theta) = \sum_{g=1}^{n} \frac{P'_g{}^2}{P_g Q_g} \quad . \qquad [2.4.3]$$

The maximum value of the information curve of a given scoring formula is obtained when the scoring weights are chosen, such that

$$w_g = \frac{P'_g}{P_g Q_g} \quad . \qquad [4.4.4]$$

So, in order to obtain the test information curve, and consequently minimize the widths of confidence bands about examinee ability estimates under the one-, two-, and three-parameter logistic models, the scoring weights should be chosen to be 1, Da_g, and $Da_g\, e^{Da_g(\theta-b_g)-\log c_g}/(1+e^{Da_g(\theta-b_g)-\log c_g})$ $(g = 1, 2, \ldots, n)$, respectively. (Information curves and the best scoring weights for other latent trait models are given by Samejima [1969, 1973].) Only for the three-parameter model are the scoring weights a function of ability level. The scoring system in the three-parameter model has the effect of reducing the weight assigned to correct answers on items where the lower asymptotes (c_g) of the item characteristic curves are large. Also, the weights for such items are smaller for low-ability examinees than for either middle- or high-ability examinees, to reflect the fact that low-ability examinees are most likely to be answering the items by guessing. For high-ability examinees, the optimum scoring weights of the items approach the quantity $Da_g (g=1, 2, \ldots, n)$.

The quantity $P'_g{}^2/P_gQ_g$ in Equation [2.5.2] is the contribution of item g to the information function of the test. For this reason it is called the *item information function.* Item information functions have an important role in determining the accuracy with which ability is estimated at different levels of θ. Each item information curve depends on the slope of the particular item characteristic curve and the conditional variance of test scores at each ability level θ. The higher the slope of the item characteristic curve and the smaller the conditional variance, the higher will be the item information curve at that particular ability level. The height of the item information curve at a particular ability level is a direct measure of the usefulness of the item for precisely measuring ability at that level.

Figures 2.4.1-2.4.6 (from Hambleton, 1979) provide three sets of typical item characteristic curves (Figures 2.4.1, 2.4.3 and 2.4.5) and corresponding item information curves (Figures 2.4.2, 2.4.4 and 2.4.6). The effects of increasing the values of the item discrimination and pseudo-chance curves are clear. High item discrimination indices result in "steeper" item characteristic curves and higher amounts of information across the ability continuum than low item discrimination indices. In addition, when item pseudo-chance level indices exceed zero, the lower asymptotes of the item characteristic curves are different from zero, and the test items provide less information, especially at the low end of the ability continuum, than test items with the pseudo-chance level values close to zero.

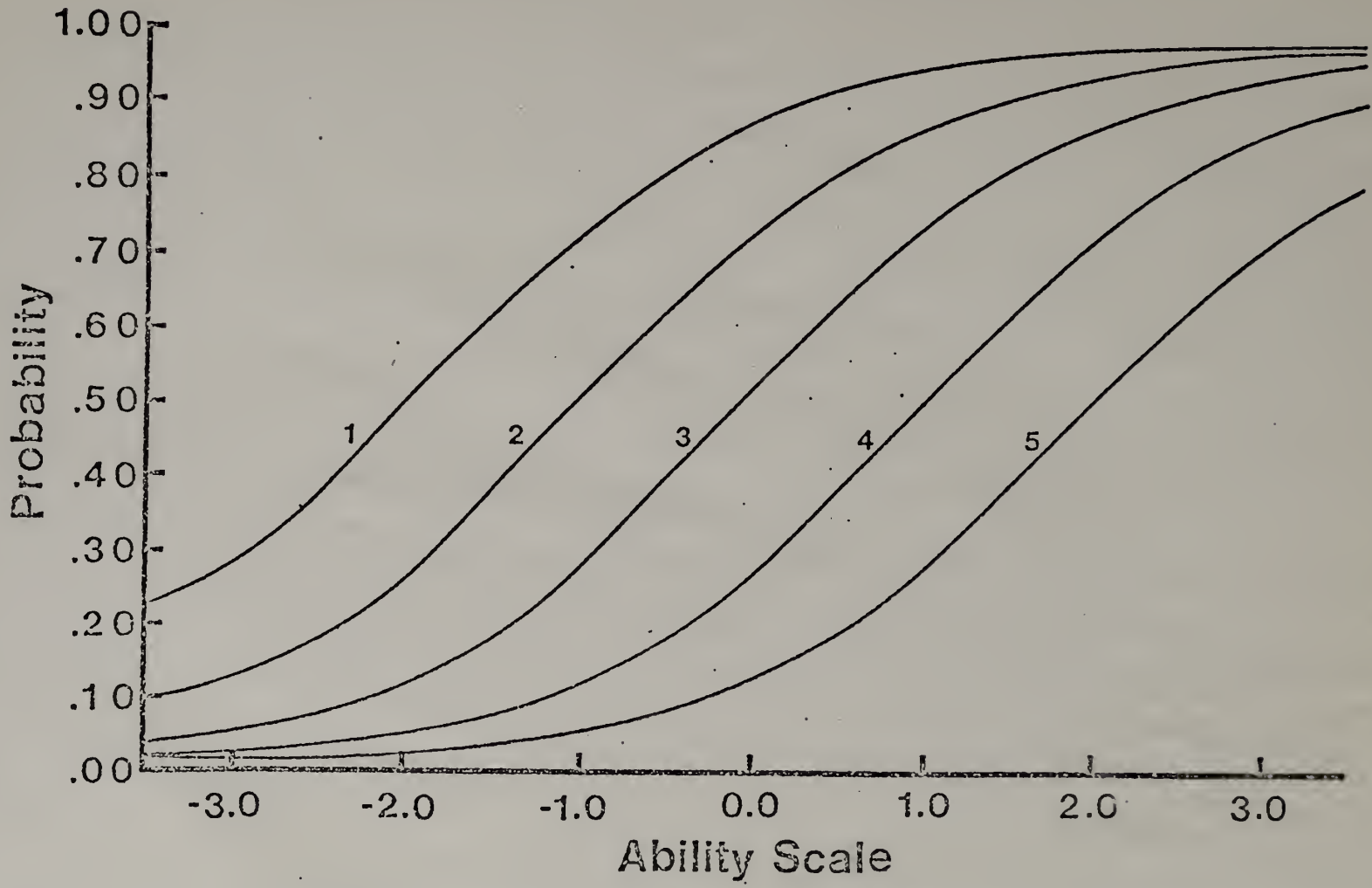


Figure 2.4.1. Graphical representation of five item characteristic curves [b=-2.0, -1.0, 0.0, 1.0, 2.0; a=.59; c=.00].

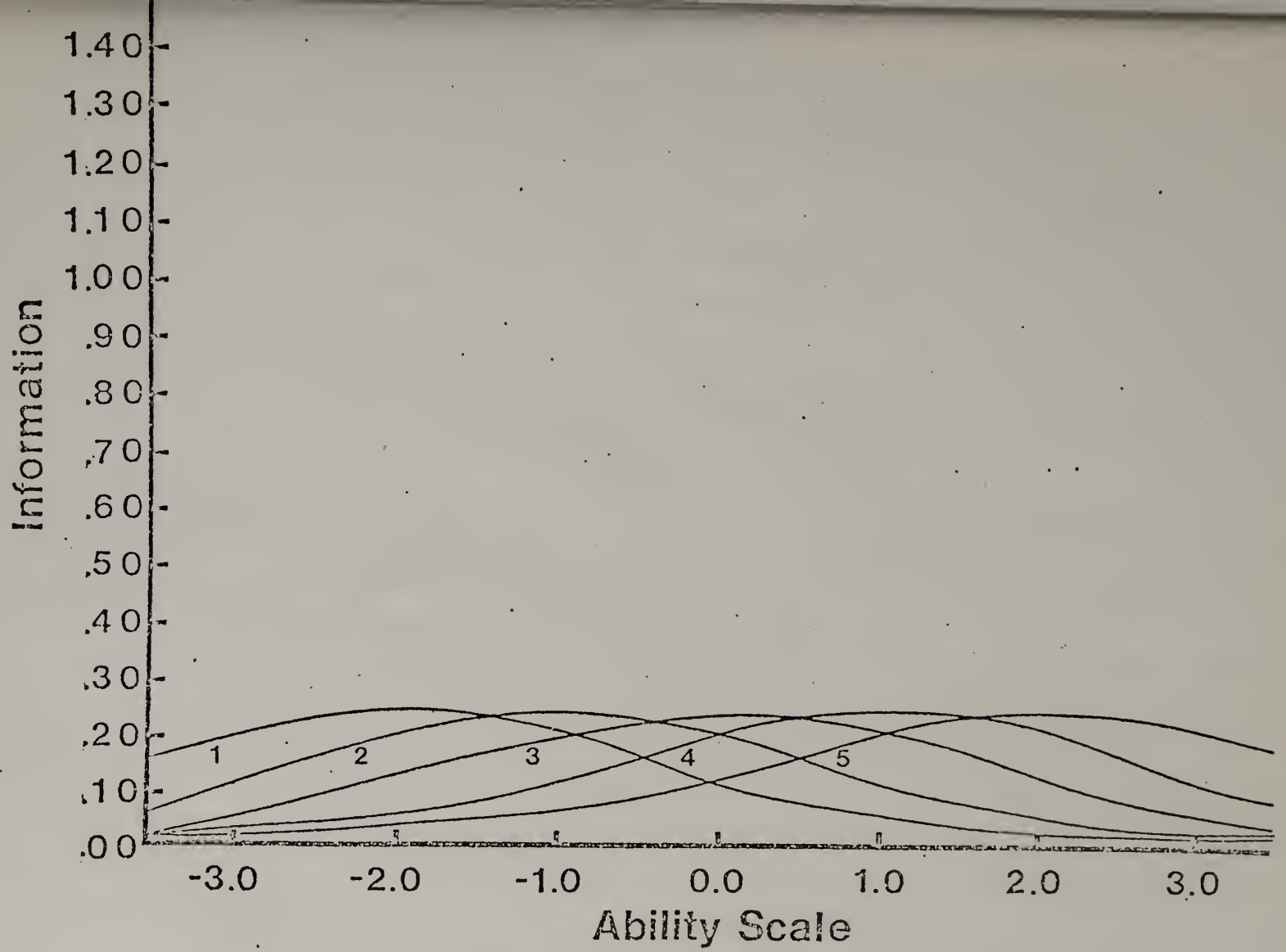


Figure 2.4.2. Graphical representation of five item information curves [b= -2.0, -1.0, 0.0, 1.0, 2.0; a= .59; c= .00].

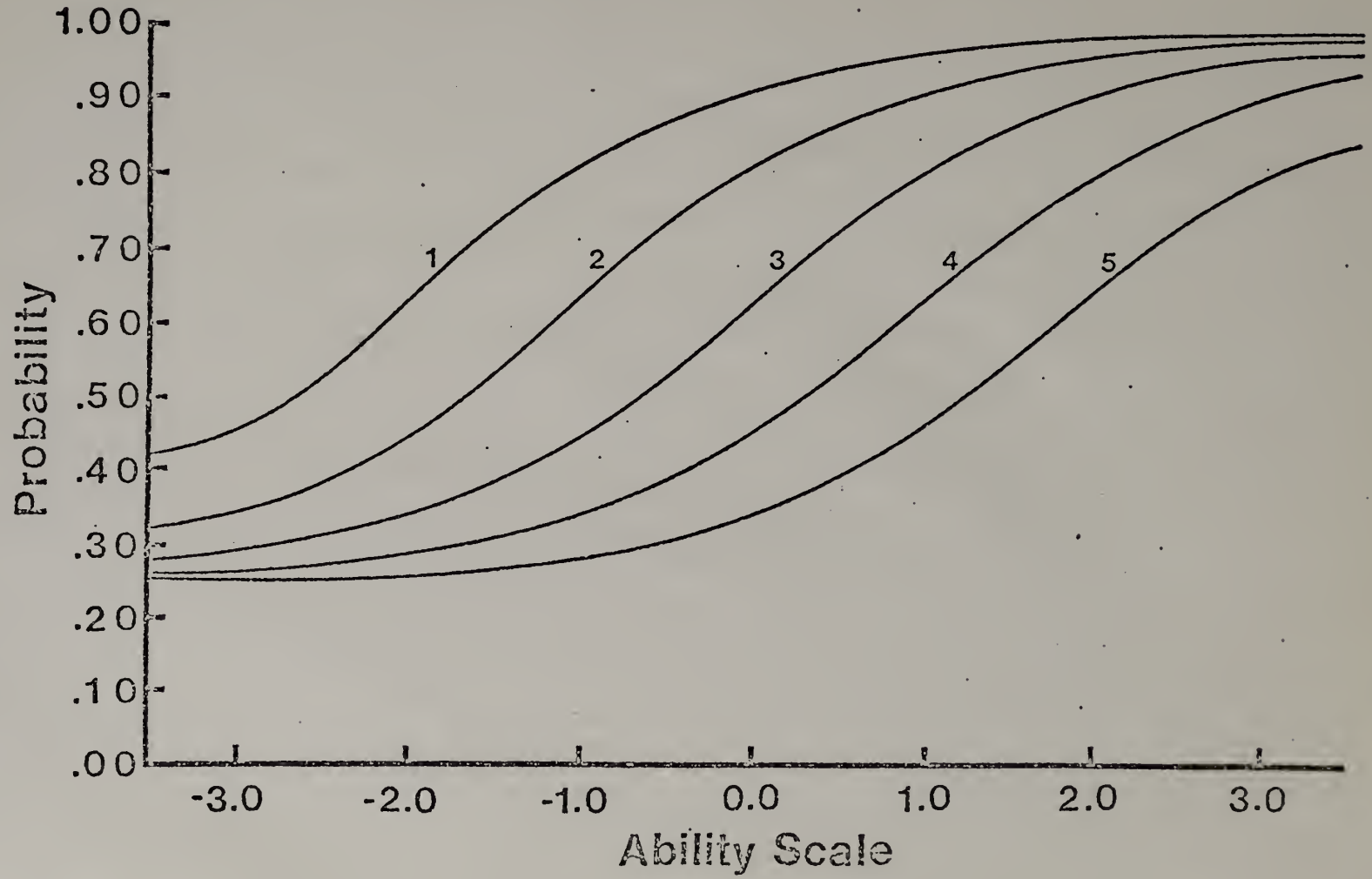


Figure 2.4.3. Graphical representation of five item characteristic curves [b= -2.0, -1.0, 0.0, 1.0, 2.0; a= .59; c= .25].

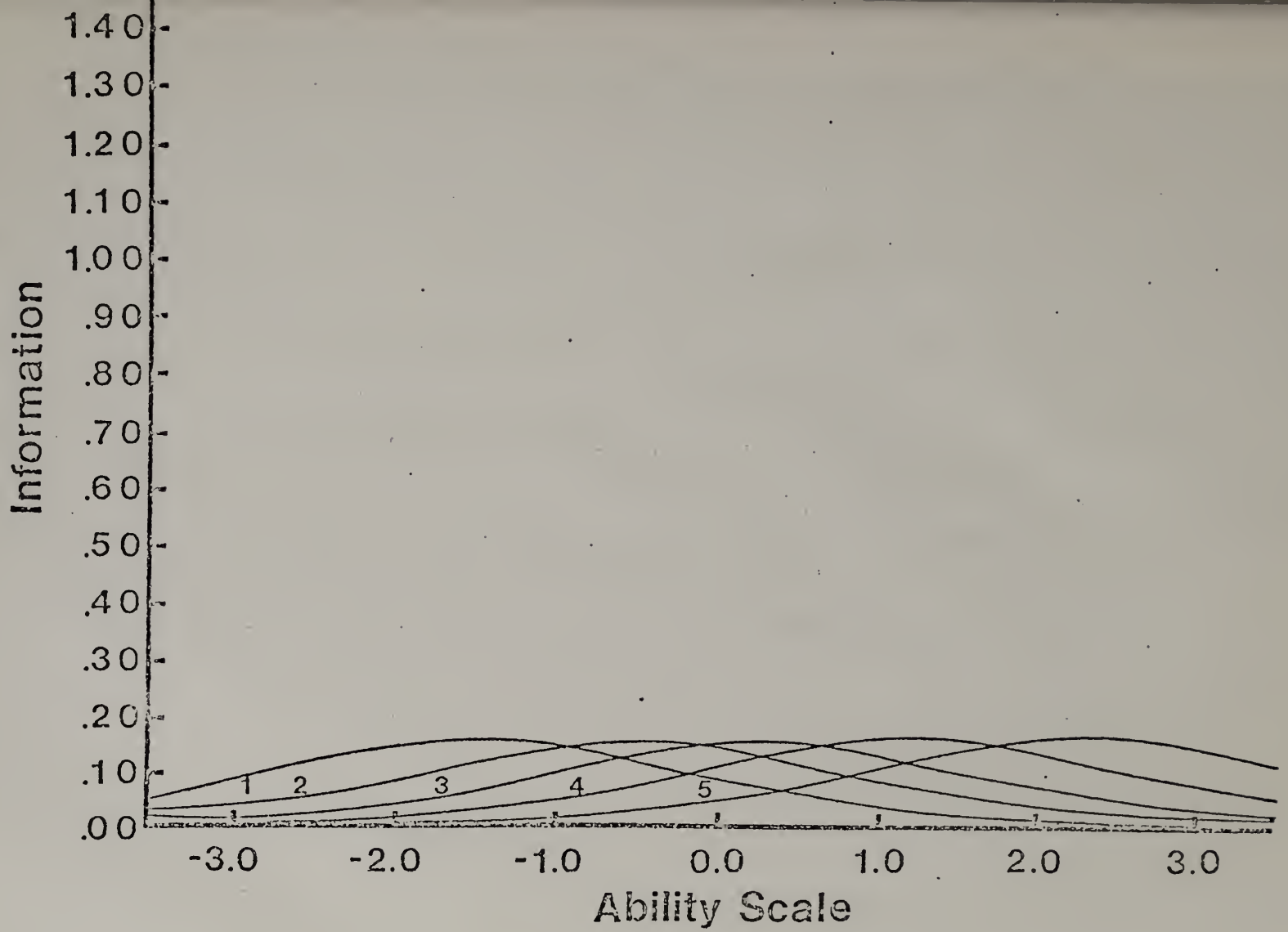


Figure 2.4.4. Graphical representation of five item information curves [b= -2.0, -1.0, 0.0, 1.0, 2.0; a= .59; c= .25].

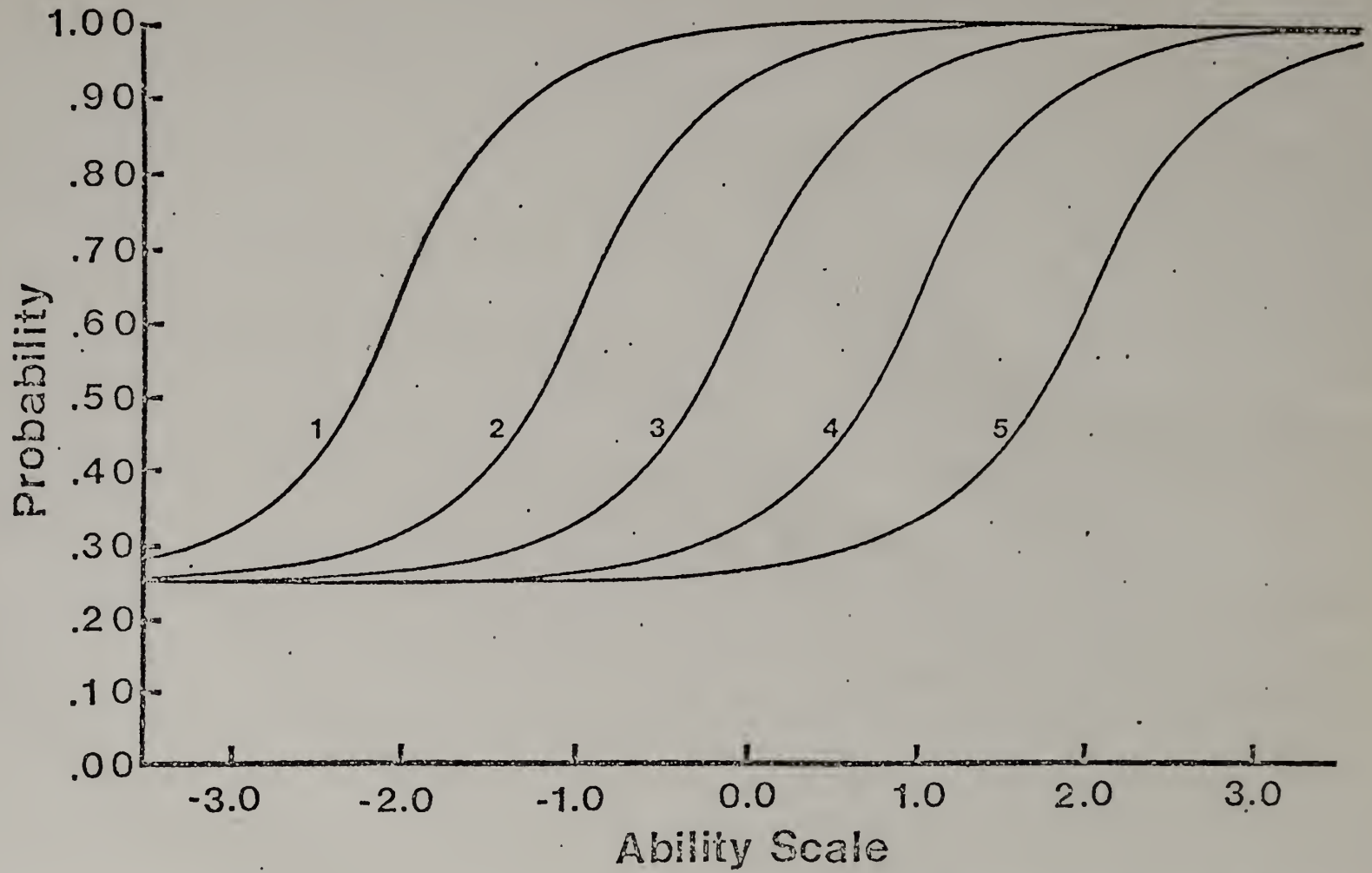


Figure 2.4.5. Graphical representation of five item characteristic curves [b=-2.0, -1.0, 0.0, 1.0, 2.0; a=1.39; c=.25].

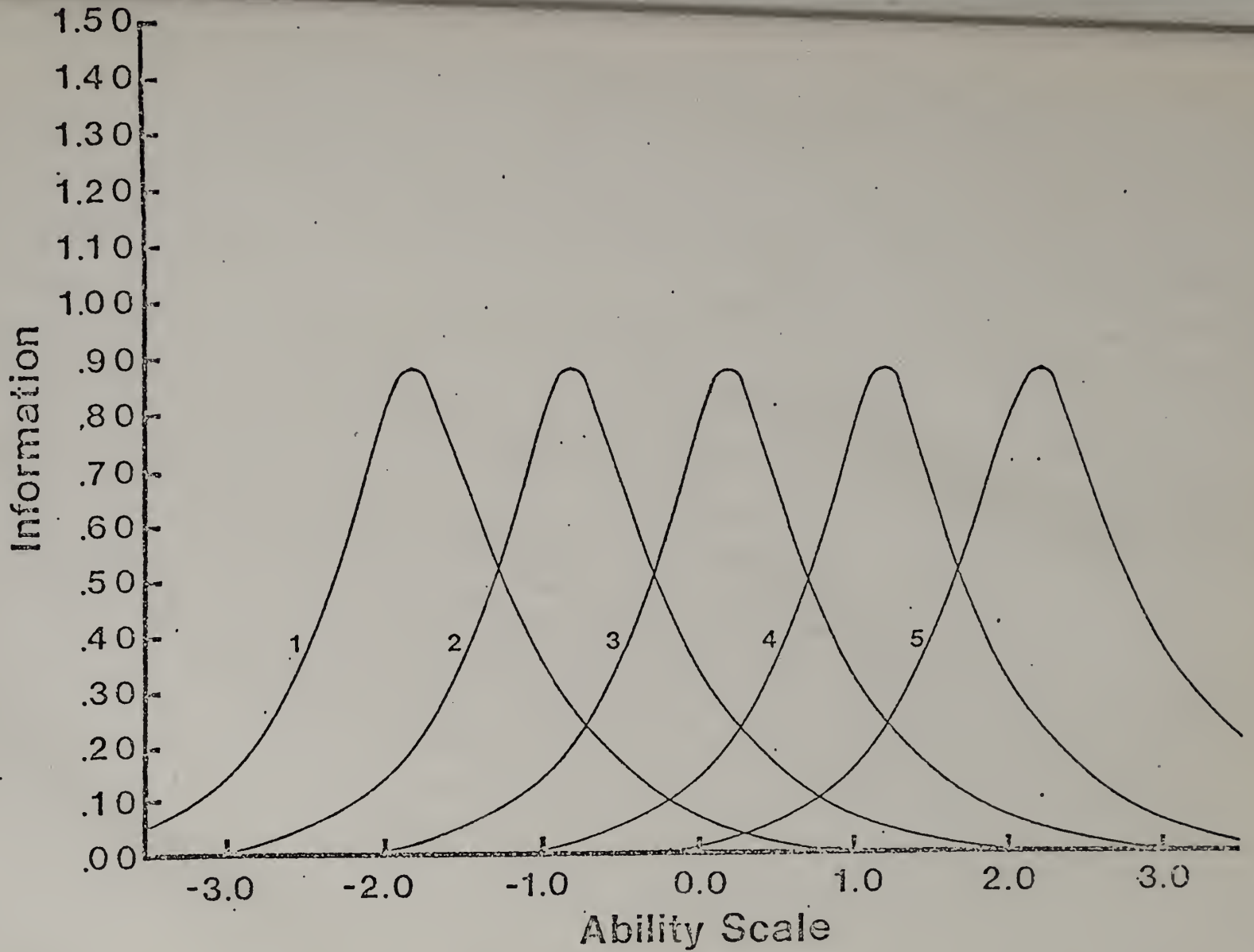


Figure 2.4.6. Graphical representation of five item information curves [b = -2.0, -1.0, 0.0, 1.0, 2.0; a = 1.39; c = .25].

From Equation [2.4.3] it is clear that items contribute independently to the test information function. Birnbaum (1968) has also shown that with his three-parameter model, an item provides maximum information at an ability level θ, where

$$\theta = b_g + \frac{1}{1.7a_g} \log_e 1/2(1 + \sqrt{1 + 8c_g}) \quad . \qquad [2.4.5]$$

If guessing is minimal, then $c_g = 0$, and $\theta = b_g$. When $c_g > 0$, the point of maximum information is shifted to the right of the item difficult value, b_g.

If non-optimal scoring weights are used with a particular item characteristic curve model, the information curve derived from Equation [2.4.1] will be lower, at all ability levels, than one that would result from the use of optimal weights. Birnbaum (1968) used the term *efficiency* to refer to the information loss due to the use of less than optimal scoring weights. Efficiency is studied by calculating the ratio of the values of the actual information curve and the test information curve at each ability level.

2.5 The Classical Test Model Versus Latent Trait Models

In view of the complexities involved in applying the latent trait models, and the restrictiveness of the assumptions underlying the models, one may ask: Why bother? After all, classical test models are well-developed, have lead to many important and useful results, and they are based on *weak* assumptions. Therefore, the classical test models can be applied to most (if not all) sets of mental test data.

In contrast, latent trait models are based on *strong* assumptions which limit their applicability to many mental test data sets. On the other hand, strong assumptions imply strong results. Perhaps the most important advantage of latent trait models (Bock & Wood, 1971) is that given a set of test items that have been fitted to a latent trait model (that is, the item parameters are known), it is possible to estimate an examinee's ability on the same ability scale from *any* subset of items in the domain of items measuring the ability. (Of course, the domain of items needs to be homogeneous in the sense of measuring a single ability. If the domain of items is too heterogeneous, the ability estimates will have little meaning.) In fact, regardless of the number of items administered, or the statistical characteristics of the items, the ability estimate for each examinee will be an unbiased estimate of true ability. Ability estimation independent of the particular choice (and number) of items represents one of the major advantages of latent trait models. Hence, latent trait models provide a way of comparing examinees even though they may have taken quite different subsets of the test items. It is this feature that makes the latent trait model most useful in the field of tailored testing. In tailored testing, examinees receive test items that are matched to their ability level. Nevertheless, the ability estimates for examinees are on a common ability scale and therefore examinees can be compared. Clearly, the usual test score metric will not permit meaningful comparisons of examinees when the tests taken by the examinees are not matched on difficulty. In latent trait models,

the difficulty of items is accounted for by the model and reflected in the ability estimates. Thus, two students, denoted A and B, receiving identical scores on an easy and difficult subset of the test items, respectively, will differ in their ability estimates (B will receive higher ability score than A).

Two other problems that can be resolved through the application of latent trait models are the problems of developing parallel-forms of a test and equating scores from one test to another that measure the same ability. Both problems can be directly resolved through fitting the test data to a latent trait model.

Another advantage of latent trait models is that the item parameters are invariant across sub-groups of examinees from the examinee population. In principle, the item parameters should remain the same regardless of the sub-group tested. Invariant item parameters have been sought by measurement specialists for a long period of time; the advantages of which are obvious for test development work. Certainly classical item statistics such as item difficulty and discrimination do not qualify. For example, it is well-known that item difficulty will vary from group to group depending upon the average ability of the group being tested.

Yet another desirable property of the latent trait models is the provision of a measure of the precision of ability estimation for each ability level. Thus, instead of a *single* estimate of the size of errors in individual examinee scores provided by the standard error of measurement, the latent trait models make it possible to

provide separate estimates of error for each examinee that are specific to each ability level.

CHAPTER III

ROBUSTNESS OF LATENT TRAIT MODELS

3.1 Introduction

While the potential usefulness of latent trait models is great, there remain many practical problems to address at the application stage. For one, how does a user go about selecting a latent trait model? One might be tempted to say that the user should always work with the more general models since these models will provide the "best" fits to the available test data. Unfortunately, the more general latent trait models (for example, the three-parameter logistic test model) require more computer time to obtain satisfactory solutions, require larger samples of examinees and longer tests, and are more difficult for practitioners to work with. Clearly, more needs to be known about the "goodness-of-fit" and "robustness" of latent trait models. Such information would aid practitioners in the important step of selecting a test model.

There has been some work on the "goodness-of-fit" between latent trait models and a variety of test data sets (see for example, Lord, 1975; Tinsley & Dawis, 1977; Wright, 1968). Hambleton, Swaminathan, Cook, Eignor, and Gifford (1978) have reviewed these as well as other studies and have noted that almost all of the studies use a chi-square statistic as the criterion measure. The problems related to using this statistic to assess goodness-of-fit will be discussed shortly.

Only one study was found in the literature that compared the fit of more than one latent trait model to the same test data sets (Hambleton & Traub, 1973). In this study, improvements were obtained in predicting test score distributions (for three tests) from the two-parameter model as compared to the one-parameter model.

On the question of model robustness (i.e., the extent to which the assumptions underlying the test model can be violated to a greater or lesser extent by the test data and be "fitted" by the model), the results of several studies have been reported (Dinero & Haertel, 1977; Hambleton, 1969; Hambleton & Traub, 1976; Panchapakesan, 1969). The results have been mixed, perhaps because of the confounding of results with sample sizes.

The problem with most of the goodness-of-fit studies and the robustness studies reported to date is that they provide no indication of the practical consequences of fitting a "less than perfect" model to a test data set. It really is of little interest to the practitioner to know that 15 out of 20 items failed to be fitted by a test model when the range of discrimination parameters reached (say) a value of .80. For one thing, if the size of the examinee sample is large enough, probably all items could be identified by a chi-square statistic of goodness-of-fit as *not* fitting the model. If the size of the examinee sample is small enough, perhaps none of the items would be misfit by the model! It is important for practitioners to see comparisons of the "fit" of latent trait models to various data sets using a criterion measure (or measures) that have some practical meaning to them. To date there have been no comparative studies of the various latent trait models using practical criteria to judge the results.

The purpose of the present research was to study, systematically, the "goodness-of-fit" of the one-, two-, and three-parameter logistic models employing a practical criterion for assessment. Using computer-simulated test data, the effects of the following four variables were studied: (1) variation in item discrimination parameters, (2) the average value of the pseudo-chance level parameters, (3) test length, and (4) the shape of the ability distribution. Artificial or similated data representing departures of varying degrees from the assumptions of the three-parameter logistic test model were generated and the "goodness-of-fit" of the three test models to the data was studied.

How should "goodness-of-fit" be measured? In some testing situations, (for example, the typical situation involving norm-referenced tests), test users desire to rank examinees based on their test score performance in a way that will closely reflect rankings based on examinee "true ability." Much effort is made by test developers to rank examinees properly (i.e., "validly") by using suitably long tests, high-quality test items, proper test conditions and so on. Therefore, a reasonable criterion for assessing the applicability of a model to a particular data set would be how effective the model is in assigning ranks to examinees that are consistent with examinee rankings based on true ability scores.

3.2 Method of Investigation

In this study, simulated data were used so that it was possible to "know" examinee ability scores. These scores served as a criterion

against which to judge the statistics derived from the three test models used to rank examinees. Three statistics, derived from the one-, two-, and three-parameter logistic models, respectively, were obtained and used to rank examinees. The rankings of examinees derived from each model (for each set of test data) were then compared to examinee "true" abilities. The Spearman rank difference formula was used to summarize the similarity between each pair of ranks (true abilities and estimates of ability from one of the models). Also reported are the average size of the discrepancies in the ranks for each group of 500 examinees.

3.2.1 Simulating the Test Data

The simulation of item response data for examinees was accomplished using the three-parameter logistic model. First, the number of examinees (N), shape of the ability distribution, and values of the ability parameters (θ_i = 1, 2, ..., N) were specified. Next, the number of items in the test (n) and values of the three item parameters (a_g, b_g, c_g, g = 1, 2, ..., n) were specified. Then the examinee and item parameters were substituted in the equation of the three-parameter logistic model to obtain a number p_{ij} ($0 \leq p_{ij} \leq 1$) representing the probability that examinee i correctly answered item j. The probabilities were arranged in a matrix P of order Nxn whose (i, j)th element was p_{ij}. P was then converted into a matrix of the item scores for examinees (1 = correct answer, 0 = incorrect answer) by comparing each p_{ij} with a random number obtained from a uniform distribution on the interval [0, 1]. If the random number was less

than or equal to p_{ij} (which would happen on the average p_{ij} of the time), p_{ij} was set equal to 1, otherwise p_{ij} was set to 0. The matrix P of zeros and ones was the simulated test data. At this point, three statistics used in estimating examinee ability were calculated:

$$\sum_{g=1}^{n} u_g, \quad \sum_{g=1}^{n} a_g u_g, \text{ and } \sum_{g=1}^{n} w_g(\theta)\, u_g ,$$

corresponding to statistics which are used in the estimation of examinee ability with the one-, two-, and three-parameter models, respectively. ($u_g = 1$ for a correst response, $u_g = 0$, otherwise.) For the three-parameter model statistic, since the item weights [$w_g(\theta)$] depend on examinee ability, three-parameter model estimates of ability were obtained for each examinee from LOGIST (Wood, Wingersky, & Lord, 1976). Once the three-parameter model estimates of ability were calculated they were used (instead of $\sum_{g=1}^{n} w_g(\theta)\, u_g$) for convenience.

The values of the examinee and item parameters were chosen as follows:

Examinee Parameters.--The number of examinees was set equal to 500. This number was sufficient to produce stable goodness-of-fit results. Two distributions of ability were considered: Uniform [-2.5, 2.5] and Normal [0, 1].

Item Parameters.--Two test lengths (20 and 40 items) were used in the simulations. Both values are typical of test lengths in common use.

In the simulation of test data, item difficulty parameters, b_g, g = 1, 2, ..., n, were selected at random from a uniform distribution on the interval [-2, 2]. An analysis of the difficulty parameters reported by Lord (1968) suggested that this decision was reasonable.

The discrimination parameters, a_g, g = 1, 2, ..., n, for the items of a simulated test were selected at random from a uniform distribution with mean = 1.12. The range of the discrimination parameters was a variable under investigation. The range was varied from 0.0 to a maximum of 1.24 [.50 to 1.74], and an intermediate value of .62 [.81 to 1.43] was also studied. The maximum range of discrimination parameters (1.24 was similar to the range of the discrimination parameters reported for the Verbal Section of the SAT (Lord, 1968).

The extent of guessing in the simulated test data was another variable under study. Two values of the average guessing parameter were considered: $\bar{c} = 0.00$, and $\bar{c} = 0.25$. All psuedo-chance level parameters were set equal to the mean value of the c-parameter under investigation. It should be noted that for all of the tests simulated in the study, it was assumed that the items were unidimensional, i.e., measured a common trait.

3.2.2 Goodness-of-Fit

The approach to goodness-of-fit was described earlier in the introductory section of this chapter. For each data set (24 in total; 2 test lengths x 2 levels of pseudo-chance parameters x 3 levels of variation in discrimination parameters x 2 ability distributions), three statistics used in estimating ability for the one-, two-, and

three-parameter models, respectively, were calculated and compared to the true ability parameters. Comparisons were made via the use of Spearman rank difference formula and the average discrepancy in ranks.

To further facilitate the interpretation of results, they are reported separately for each half of the ability distribution as well as for the total ability distribution.

3.3 Results

The results of the computer simulations are summarized in Tables 3.3.1 to 3.3.6. The first row of each table was inserted to serve as a check on the calculations.

For convenience the results will be discussed in point form around the variables under study:

Level of Variation in Discrimination Parameters

1. For the values studied, using discrimination parameters as item weights contributed very little to the correct ranking of examinees.

Level of Pseudo-Chance Level Parameters

2. With the twenty-item tests, the three-parameter model was considerably more effective at ranking examinees correctly in the lower half of the ability distribution. Correlations were about .08 higher ($\sim$.75 to $\sim$.83) in the uniform distribution of ability and about .08 higher in the normal distribution ($\sim$.65 to $\sim$.73). The improvement in the average absolute difference in rank order was about 13.

3. With the forty-item tests, the three-parameter model was also somewhat more effective at ranking examinees correctly in the lower half of the ability distribution. Correlations were about .04 higher in both ability distributions. The improvement in the average absolute difference in rank order was about 8. The reduction in effectiveness of the

Table 3.3.1

Summary of the Goodness-of-Fit Results
(Uniform Ability Distribution,[1] θ = -2.5 to 0.0)

Test Length	Variation in Discrimination Parameters	Pseudo-Chance Level Parameters	Test Score Statistics $\bar{X}$	SD	Comparison of Estimates: True Versus One Parameter Model r^2	AAD[3]	True Versus Two Parameter Model r	AAD	True Versus Three Parameter Model r	AAD
20	0.00	.00	5.03	3.00	.881	54.238	.881	54.238	.881	54.238
20	0.00	.25	8.98	2.86	.765	76.610	.765	76.610	.827	64.984
20	.81 to 1.43	.00	5.24	3.10	.877	56.068	.876	56.406	.876	56.404
20	.81 to 1.43	.25	9.01	2.84	.760	77.144	.764	76.900	.833	64.284
20	.50 to 1.74	.00	5.36	3.02	.874	56.496	.874	56.558	.874	56.562
20	.50 to 1.74	.25	9.12	2.83	.747	80.076	.750	79.920	.827	65.770
40	0.00	.00	9.58	6.22	.944	36.482	.944	36.482	.944	36.482
40	0.00	.25	17.82	5.33	.868	58.578	.868	58.578	.908	48.704
40	.81 to 1.43	.00	10.14	6.37	.949	36.504	.949	36.474	.949	36.474
40	.81 to 1.43	.25	17.98	5.41	.872	57.662	.875	56.860	.912	48.014
40	.50 to 1.74	.00	9.97	6.39	.942	37.862	.946	36.962	.946	36.742
40	.50 to 1.74	.25	18.18	5.41	.870	57.824	.876	56.872	.910	48.222

[1]N = 500
[2]Spearman Rank-Difference Formula
[3]Average absolute difference in rank order

Table 3.3.2

Summary of the Goodness-of-Fit Results
(Uniform Ability Distribution,[1] θ = 0.00 to +2.5)

Test Length	Variation in Discrimination Parameters	Pseudo-Chance Level Parameters	Test Score Statistics $\bar{X}$	SD	Comparison of Estimates: True Versus One Parameter Model r^2	AAD[3]	True Versus Two Parameter Model r	AAD	True Versus Three Parameter Model r	AAD
20	0.00	.00	14.99	2.82	.883	54.450	.877	55.624	.877	55.624
20	0.00	.25	16.21	2.13	.835	63.676	.828	65.350	.829	65.726
20	.81 to 1.43	.00	15.12	2.75	.891	52.234	.881	55.376	.881	55.382
20	.81 to 1.43	.25	16.16	2.14	.847	63.802	.832	65.018	.841	63.190
20	.50 to 1.74	.00	14.93	2.79	.872	56.988	.882	55.384	.882	55.470
20	.50 to 1.74	.25	16.36	2.09	.797	71.570	.797	70.720	.804	69.164
40	0.00	.00	31.73	5.55	.940	39.034	.936	40.496	.936	40.496
40	0.00	.25	33.52	4.37	.903	50.188	.898	51.046	.896	50.852
40	.81 to 1.43	.00	31.30	5.53	.935	40.648	.932	41.832	.932	41.848
40	.81 to 1.43	.25	33.47	4.26	.908	49.142	.903	50.554	.905	50.266
40	.50 to 1.74	.00	31.15	5.39	.934	40.788	.939	38.932	.939	38.940
40	.50 to 1.74	.25	33.40	4.16	.890	52.882	.892	52.898	.893	52.678

[1]N = 500
[2]Spearman Rank-Difference Formula
[3]Average absolute difference in rank order

Table 3.3.3

Summary of the Goodness-of-Fit Results
(Uniform Ability Distribution,[1] θ = - 2.5 to +2.5)

Test Length	Variation in Discrimination Parameters	Pseudo-Chance Level Parameters	Test Score Statistics $\bar{X}$	SD	Comparison of Estimates: True Versus One Parameter Model r^2	AAD[3]	True Versus Two Parameter Model r	AAD	True Versus Three Parameter Model r	AAD
20	0.00	.00	9.91	5.84	.970	28.264	.970	28.368	.970	28.368
20	0.00	.25	12.40	4.43	.932	41.850	.931	41.972	.949	36.968
20	.81 to 1.43	.00	9.97	5.63	.969	28.808	.969	29.138	.969	29.140
20	.81 to 1.43	.25	12.28	4.35	.931	42.402	.928	43.932	.943	38.594
20	.50 to 1.74	.00	10.50	5.58	.965	30.826	.966	30.140	.966	30.140
20	.50 to 1.74	.25	12.40	4.54	.932	42.200	.931	42.726	.942	39.016
40	0.00	.00	20.99	12.21	.984	20.438	.984	20.614	.984	20.614
40	0.00	.25	24.54	9.40	.964	30.130	.964	30.260	.971	27.018
40	.81 to 1.43	.00	20.31	12.54	.983	21.088	.983	21.250	.983	21.254
40	.81 to 1.43	.25	24.58	9.36	.962	30.690	.962	30.750	.971	27.738
40	.50 to 1.74	.00	19.93	12.12	.981	22.478	.982	21.814	.982	21.808
40	.50 to 1.74	.25	24.94	9.16	.962	31.490	.964	30.498	.972	27.302

[1]N = 500
[2]Spearman Rank-Difference Formula
[3]Average absolute difference in rank order

Table 3.3.4

Summary of the Goodness-of-Fit Results
(Lower Half of Normal Ability Distribution,[1] $\bar{X}_\theta = 0.00$, $SD_\theta = 1.00$)

Test Length	Variation in Discrimination Parameters	Pseudo-Chance Level Parameters	Test Score Statistics $\bar{X}$	Test Score Statistics SD	Comparison of Estimates: True Versus One Parameter Model r^2	True Versus One Parameter Model AAD[3]	True Versus Two Parameter Model r	True Versus Two Parameter Model AAD	True Versus Three Parameter Model r	True Versus Three Parameter Model AAD
20	0.00	.00	6.77	2.69	.817	65.584	.817	65.584	.817	65.584
20	0.00	.25	10.04	2.54	.649	94.928	.649	94.928	.736	82.536
20	.81 to 1.43	.00	6.72	2.66	.835	62.716	.830	63.262	.830	63.312
20	.81 to 1.43	.25	10.10	2.56	.653	95.184	.645	95.774	.729	83.486
20	.50 to 1.74	.00	7.05	2.61	.796	70.646	.801	69.428	.801	69.414
20	.50 to 1.74	.25	10.25	2.57	.655	94.628	.641	95.800	.725	83.380
40	0.00	.00	13.61	5.48	.909	46.026	.909	46.026	.909	46.026
40	0.00	.25	20.06	4.78	.813	68.700	.813	68.700	.848	61.626
40	.81 to 1.43	.00	13.65	5.55	.903	48.234	.908	47.276	.907	47.280
40	.81 to 1.43	.25	20.19	4.86	.810	68.078	.816	67.048	.852	60.094
40	.50 to 1.74	.00	14.29	5.78	.901	48.218	.909	46.580	.909	46.582
40	.50 to 1.74	.25	20.47	4.90	.805	69.010	.813	68.662	.848	61.578

[1] $N = 500$
[2] Spearman Rank-Difference Formula
[3] Average absolute difference in rank order

Table 3.3.5

Summary of the Goodness-of-Fit Results
(Upper Half of Normal Ability Distribution,[1] $\bar{X}_\theta = 0.00$, $SD_\theta = 1.00$)

Test Length	Variation in Discrimination Parameters	Pseudo-Chance Level Parameters	Test Score Statistics $\bar{X}$	Test Score Statistics SD	Comparison of Estimates: True Versus One Parameter Model r^2	True Versus One Parameter Model AAD[3]	True Versus Two Parameter Model r	True Versus Two Parameter Model AAD	True Versus Three Parameter Model r	True Versus Three Parameter Model AAD
20	0.00	.00	13.37	2.62	.844	60.506	.844	60.808	.844	60.808
20	0.00	.25	15.12	2.20	.761	75.752	.759	76.158	.769	75.076
20	.81 to 1.43	.00	13.37	2.61	.853	61.088	.852	61.596	.852	61.606
20	.81 to 1.43	.25	15.12	2.18	.759	76.406	.757	78.024	.769	75.628
20	.50 to 1.74	.00	13.43	2.52	.834	64.792	.846	63.084	.846	63.076
20	.50 to 1.74	.25	15.11	2.12	.749	78.686	.752	79.920	.767	77.012
40	0.00	.00	27.96	4.93	.895	50.714	.895	50.748	.895	50.748
40	0.00	.25	31.02	3.75	.823	65.180	.822	65.448	.833	64.236
40	.81 to 1.43	.00	28.28	4.91	.894	51.252	.898	50.212	.898	50.226
40	.81 to 1.43	.25	31.11	3.81	.824	65.924	.830	64.838	.839	63.160
40	.50 to 1.74	.00	28.39	4.90	.892	51.014	.898	49.954	.898	49.952
40	.50 to 1.74	.25	31.20	3.77	.808	67.604	.822	64.512	.828	63.958

[1] N = 500
[2] Spearman Rank-Difference Formula
[3] Average absolute difference in rank order

Table 3.3.6

Summary of the Goodness-of-Fit Results
(Normal Ability Distribution,[1] $\bar{X}_\theta = 0.0$, $SD_\theta = 1.0$)

Test Length	Variation in Discrimination Parameters	Pseudo-Chance Level Parameters	Test Score Statistics $\bar{X}$	Test Score Statistics SD	Comparison of Estimates: True Versus One Parameter Model r^2	True Versus One Parameter Model AAD[3]	True Versus Two Parameter Model r	True Versus Two Parameter Model AAD	True Versus Three Parameter Model r	True Versus Three Parameter Model AAD
20	0.00	.00	10.30	4.27	.940	36.844	.940	36.906	.940	36.906
20	0.00	.25	12.37	3.49	.883	53.940	.883	53.896	.908	47.554
20	.81 to 1.43	.00	10.43	4.33	.943	35.868	.944	35.988	.944	35.982
20	.81 to 1.43	.25	12.40	3.46	.882	54.306	.883	54.336	.905	48.610
20	.50 to 1.74	.00	10.51	4.20	.930	41.114	.932	40.958	.932	40.962
20	.50 to 1.74	.25	12.48	3.50	.873	55.726	.865	57.942	.881	53.128
40	0.00	.00	21.22	9.21	.971	26.598	.971	26.620	.971	26.620
40	0.00	.25	25.78	7.11	.946	36.442	.946	36.464	.956	33.030
40	.81 to 1.43	.00	20.90	9.39	.973	25.196	.973	25.536	.973	25.534
40	.81 to 1.43	.25	25.88	7.01	.939	38.864	.942	37.648	.952	34.148
40	.50 to 1.74	.00	20.87	8.99	.970	27.038	.972	25.878	.972	25.874
40	.50 to 1.74	.25	25.91	6.99	.937	38.794	.941	37.330	.951	34.676

[1] $N = 500$
[2] Spearman Rank-Difference Formula
[3] Average absolute difference in rank order

three-parameter model weights was to be expected with the longer tests. Gulliksen (1950) noted the insignificance of scoring weights when the test gets longer and test items are positively correlated.

4. For examinees in the upper half of the ability distribution, and for the data sets studied, the number rights score was about an effective as the more complicated scoring weights used in the two- and three-parameter models.

Shape of the Ability Distribution

5. As expected, correlations tended to be higher for the uniformly distributed ability scores.

Test Length

6. It is interesting to observe the increases in correlations due to doubling the length of the test. Again, as expected they tended to be rather small.

3.4 Conclusions

From the data sets analyzed in this study, it is clear that there are some sizable gains to be expected with modest length tests (n = 20) in the correct ordering of examinees at the lower end of the ability continuum when three-parameter model estimates are used (as opposed to the number right score). The gains were cut roughly in half when the tests were doubled (n = 40) in length. It was surprising that item discrimination parameters as weights had so little effect on the results. On the other hand, Gulliksen (1950) summarized the research on item weights nearly thirty years ago and came to essentially the same conclusion! This emphasizes an important point. To the extent that the simulated data sets are typical of real data, it would appear that the application of latent trait models to the problem of "ranking" examinees is probably not worth the trouble except in those situations

where gains of the size noted for lower ability examinees are important. The number right score does nearly as good a job of ranking examinees as the most complicated scoring methods.

The usual cautions that apply to generalizing the results from a single study must be made here. For one, it is possible that the simulations do not closely reflect real data. Second, the criterion measure of goodness-of-fit seems suitable for the situation in which a user desires to make norm-referenced interpretations of his/her test scores. There are many other test situations (for example, those involving tailored tests, test score equating, and criterion-referenced tests) where a different criterion to judge the quality of a solution would be more suitable. Third, the results of the study provide a somewhat unfair comparison of the two-parameter model with the other two models. This is because the item discrimination parameters used in the weighting process to derive statistics for ability estimation would have been somewhat different had the "best-fitting" two-parameter curves to the three-parameter item characteristic curves been used. The item discrimination parameters in the "best fitting" two-parameter curves would have differed somewhat from those defined in the three-parameter curves they were fitted to.

A final point should also be stressed. The correlation results of the one-parameter model and (to a much lesser extent) the two-parameter model are inflated (to an unknown extent) because of tied scores. Therefore, the true differences in the reported correlations are somewhat larger than those reported in Tables 3.3.1 to 3.3.6.

CHAPTER IV

EFFECTS OF TEST LENGTH AND SAMPLE SIZE ON THE ESTIMATES OF PRECISION OF LATENT ABILITY SCORES

4.1 Introduction

One of the features of using any latent trait model is the possibility of specifying a "target information curve" and then selecting test items from an item pool to produce a test with the features characterized by the "target information curve." A target information curve describes the desired level of "information" at each point on the ability scale underlying examinee test performance. Information, in turn, is directly related to the degree of precision of ability estimates at different points on the ability continuum. In fact, as long as a test is *not* too short, the standard error of estimation at a particular ability level is equal to one divided by the square root of information provided by the test at the ability level in question ($SEE(\theta) = 1/\sqrt{\text{information}(\theta)}$). In practice, since the contribution of each test item to the test information curve (referred to as a "score information curve" when item parameter estimates are used instead of the item parameter values) is known (once the item parameter values or the item parameter estimates are specified), it is possible to select test items from a pool of "calibrated" test items (i.e., a pool of test items with associated

parameter estimates) to produce a "score information curve" which approximates a desired "target information curve."

One of the problems with the paradigm offered above for test development is the imprecision associated with the item parameter estimates. Score information curves (and therefore the associated standard errors of ability estimates) will depend on the precision of item parameter estimates. In turn, precision of item parameter estimates is influenced by the examinee sample size used to estimate the item parameters, and by the length of the test. This study was designed to address two practical questions which are of some importance and interest to test developers:

1. What are the effects of examinee sample size and test length on the precision of standard error of ability estimation curves?

2. What effects do the statistical characteristics of an item pool have on the precision of standard error of ability estimation curves?

A computer simulation study was chosen as the mode of investigation for the two questions because of the large number of variables which were to be studied, and the need to "know" in some instances, the values of the item parameters.

The remainder of this chapter is divided into three sections: (1) Method of Investigation, (2) Results, and (3) Conclusion.

4.2 Method of Investigation

4.2.1 Description of the Variables

(a) Test Length. Tests of three lengths were considered: 10, 20, and 80 items. These test lengths were chosen to represent:

(1) tests that are about as short as any that are used in practice (n=10), (2) intermediate length tests (n=20), and (3) tests that represent the typical length of most long tests that are encountered in practical testing situations (n=80).

(b) Ability Distribution. In this particular study, ability scores were simulated to be normally distributed (mean = 0, sd = 1). The data was simulated to conform with a basic assumption made in the item parameter estimation method selected for the study (Urry, 1974). Urry's method was chosen for the study because (1) the method has been extensively used and found to give acceptable results and (2) Urry's computer program is relatively inexpensive as compared to the alternative program (LOGIST) used in the previously reported study.

(c) Sample Size. Three examinee sample sizes were chosen: 50, 200, and 1000. The smallest sample size (N=50) is considerably smaller than anyone should use in practice. It was chosen to identify the "worst possible" results that could be expected. The other two sample sizes define minimum and maximum sample sizes typically used in test development work with latent trait models.

(d) Item Pools. Ranges of parameter values for items in the two pools are shown below:

Item Parameter	Range of Values Pool One	Pool Two
Difficulty (b)	-2.00 to 2.00	-1.00 to 1.00
Discrimination (a)	.60 to 2.00	.60 to 1.50
Pseudo-Chance (c)	.25 to .25	.25 to .25

The differences between the two item pools can be described as follows: Items in pool one had a wider range of difficulty and discrimination values.

4.2.2 Simulation of Data

The eight steps in the simulation study were as follows:

1. Item pool one was selected for study.

2. A test length (10, 20, or 80 items) and a sample size (50, 200, or 1000 examinees) were selected. A sample of examinee ability scores was drawn from a normal distribution (mean=0, sd=1).

3. Using a computer program, DATAGEN (Hambleton & Rovinelli, 1973), (1) item parameters, given the constraints of the item pool under investigation, and (2) examinee item scores were produced. The computer program assumed the correctness of the three-parameter logistic model, used the ability scores from step 2 and item parameters generated at this step, to produce probabilities of correct answers for examinees to the test items. These probabilities, in turn, were converted to examinee item scores (0 or 1) via the use of a random number generator.

4. The examinee item scores from step 3 were used in Urry's computer program to estimate item and ability parameters. However, only the item parameter estimates were used further in this particular study.

5. The item parameter estimates were used in Equation 2 to obtain SEE (θ). The value of SEE (θ) at seven ability levels (θ = -3.00, -2.00, -1.00, 0.00, 1.00, 2.00, 3.00) was calculated.

6. Steps 3 to 5 were repeated three times to obtain three estimates of SEE (θ). All item and ability parameter values for the three runs were identical. The particular examinee item scores varied from one run to the next because of the probabilistic nature of the score outcomes.

7. Steps 3 to 6 were repeated for each combination of test length and sample size (3x3=9).

8. Steps 2 to 7 were repeated with the second item pool. In all, 54 sets of test data were considered in the study.

4.3 Results and Discussion

4.3.1 Effects of Sample Size and Test Length of the Precision of Standard Error of Ability Estimation Curves

Tables 4.3.1 to 4.3.6 contain the SEE Curves with Item Pool One obtained for three replications of three examinees sample sizes (N=50, 200, 1000) and three test lengths (n=10, 20, 80) and reported for seven ability levels. Tables 4.3.1 to 4.3.3 and 4.3.4 to 4.3.6 contain the same information. What differs is the way the data are organized in the two sets of Tables. Data have been arranged in Tables 4.3.1 to 4.3.3 to facilitate an examination of the effect of sample size on SEE Curves. The data presented in Tables 4.3.4 to 4.3.6 have been arranged to facilitate an examination of the effect of test length on SEE Curves. Test lengths and sample sizes given under the column headed "actual" are the number of items and examinees remaining after a satisfactory set of item and ability parameter estimates are obtained from Urry's computer program.

For ease of interpretation, the same data reported in Tables 4.3.1 to 4.3.6 is presented in graphical form in Figure 4.3.1.

Tables 4.3.7 to 4.3.12 contain similar data to Tables 4.3.1 to 4.3.6. Tables 4.3.7 to 4.3.12 contain SEE Curves obtained with Item Pool Two. (There is no figure, however, corresponding to Figure 4.3.1 for Item Pool Two.) Tables 4.3.13 and 4.3.14 were constructed to organize the data reported in Tables 4.3.1 to 4.3.12 to facilitate the interpretation of results.

Table 4.3.1

Summary of Standard Error Estimates[1] for Various Sample Sizes and Ability Levels with a Heterogeneous Item Pool (Test Length = 10 Items)

Sample Size	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
50	1	10	34	.66	.33	.67	.22	.75	1.60	2.19
	2	10	34	2.40	1.88	.56	1.04	.20	1.34	1.37
	3	9	34	.73	.57	1.03	.22	.58	.43	2.19
200	1	10	172	.64	.21	.52	2.15	1.60	1.50	1.48
	2	10	137	.22	.51	.36	1.30	.37	.96	2.45
	3	10	174	2.63	2.14	.27	2.75	.92	.76	1.91
1000	1	10	841	.98	.26	.58	1.43	3.33	.57	1.18
	2	10	833	1.03	1.03	.67	1.05	.45	1.01	1.06
	3	10	892	2.44	.49	.67	.30	.29	.89	1.33

[1]All estimates have been adjusted to correspond to 10-item tests.

Table 4.3.2

Summary of Standard Error Estimates for Various Sample Sizes and Ability Levels with a Heterogeneous Item Pool (Test Length = 20 Items)

Sample Size	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
50	1	20	50	2.84	.70	.35	.30	.31	.44	1.23
	2	20	50	1.93	1.53	.39	.32	.24	.45	1.19
	3	20	46	2.07	.83	.58	.31	.36	.68	1.48
200	1	20	193	--	.57	.26	.39	.33	.50	.77
	2	20	196	--	1.51	.37	.34	.25	.53	.86
	3	20	196	--	1.03	.22	.49	.34	.40	1.15
1000	1	20	955	--	1.05	.48	.33	.33	.45	.82
	2	20	969	--	1.18	.37	.33	.37	.40	.99
	3	20	968	--	1.56	.40	.42	.32	.43	1.07

Table 4.3.3

Summary of Standard Error Estimates[1] for Various Sample Sizes and Ability Levels with a Heterogeneous Item Pool (Test Length = 80 Items)

Sample Size	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
50	1	74	50	1.10	.35	.14	.14	.24	.24	.45
	2	79	50	1.06	.48	.25	.17	.13	.32	.49
	3	77	50	.93	.20	.19	.15	.17	.29	.48
200	1	80	200	.89	.26	.22	.24	.19	.25	.44
	2	80	200	.62	.29	.25	.19	.21	.25	.46
	3	80	200	1.06	.35	.21	.19	.20	.25	.48
1000	1	80	999	1.00	.35	.23	.21	.21	.24	.40
	2	80	1000	.98	.32	.23	.22	.21	.23	.43
	3	80	1000	1.08	.34	.20	.21	.20	.24	.46

[1]All estimates have been adjusted to correspond to 80-item tests.

Table 4.3.4

Summary of Standard Error Estimates for Various Test Lengths and Ability Levels with a Heterogeneous Item Pool (Sample Size = 50 Examinees)

Test Length	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
10	1	10	34	.66	.33	.67	.22	.75	1.60	2.19
	2	10	34	2.40	1.88	.56	1.04	.20	1.34	1.37
	3	9	34	.73	.57	1.03	.22	.58	.43	2.19
20	1	20	50	2.84	.70	.35	.30	.31	.44	1.23
	2	20	50	1.93	1.53	.39	.32	.24	.45	1.19
	3	20	46	2.07	.83	.58	.31	.36	.68	1.48
80	1	74	50	1.10	.35	.14	.14	.24	.24	.45
	2	79	50	1.06	.48	.25	.17	.13	.32	.49
	3	77	50	.93	.20	.19	.15	.17	.29	.48

Table 4.3.5

Summary of Standard Error Estimates for Various Test Lengths and Ability Levels with a Heterogeneous Item Pool (Sample Size = 200 Examinees)

Test Length	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
10	1	10	172	.64	.21	.52	2.15	1.60	1.50	1.48
	2	10	137	.22	.51	.36	1.30	.37	.96	2.45
	3	10	174	2.63	2.14	.27	2.75	.92	.76	1.91
20	1	20	193	--	.57	.26	.39	.33	.50	.77
	2	20	196	--	1.51	.37	.34	.25	.53	.86
	3	20	196	--	1.03	.22	.49	.34	.40	1.15
80	1	80	200	.89	.26	.22	.24	.19	.25	.44
	2	80	200	.62	.29	.25	.19	.21	.25	.46
	3	80	200	1.06	.35	.21	.19	.20	.25	.48

Table 4.3.6

Summary of Standard Error Estimates for Various Test Lengths and Ability Levels with a Heterogeneous Item Pool (Sample Size = 1000 Examinees)

Test Length	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
10	1	10	841	.98	.26	.58	1.43	3.33	.57	1.18
	2	10	833	1.03	1.03	.67	1.05	.45	1.01	1.06
	3	10	892	2.44	.49	.67	.30	.29	.89	1.33
20	1	20	955	--	1.05	.48	.33	.33	.45	.82
	2	20	969	--	1.18	.37	.33	.37	.40	.99
	3	20	968	--	1.56	.40	.42	.32	.43	1.07
80	1	80	999	1.00	.35	.23	.21	.21	.24	.40
	2	80	1000	.98	.32	.23	.22	.21	.23	.43
	3	80	1000	1.08	.34	.20	.21	.20	.24	.46

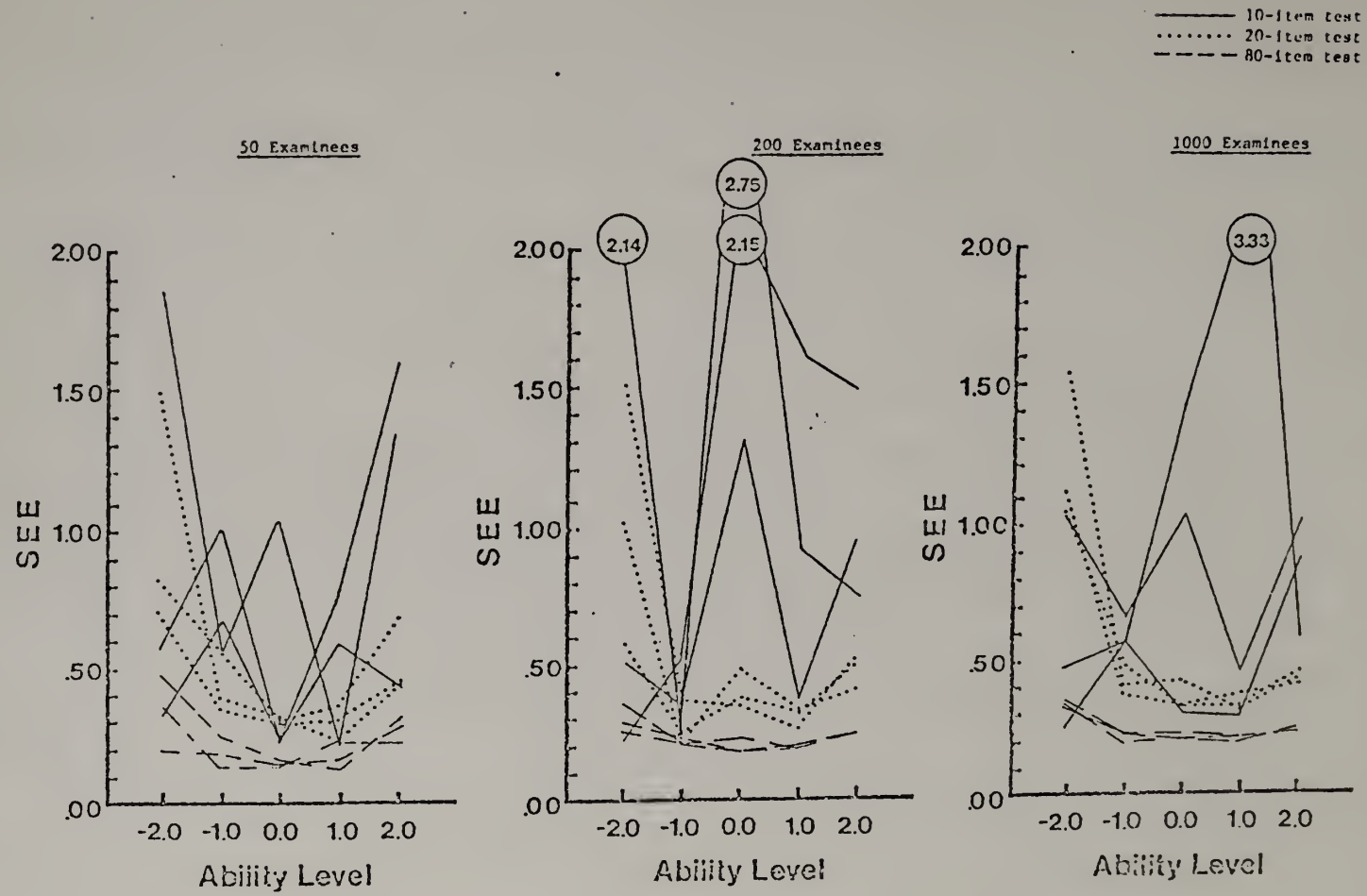


Figure 4.3.1. Standard errors of estimation associated with three test lengths (10, 20, and 80 test items) at five ability levels and reported for three sample sizes (50, 200, and 1000 examinees). (Each combination of conditions was replicated three times.)

Table 4.3.7

Summary of Standard Error Estimates for Various Sample Sizes and Ability Levels with a Homogeneous Item Pool (Test Length = 10 Items)

Sample Size	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
50	1	10	37	--	.66	.53	.51	1.43	1.05	2.71
	2	10	48	--	--	.79	.24	.48	1.59	--
	3	10	45	--	.80	.39	.42	.70	1.67	4.02
200	1	10	185	--	3.13	.41	.48	.41	1.16	4.03
	2	10	192	--	.52	.40	.65	.39	1.20	4.44
	3	10	179	--	3.89	.35	.60	.46	1.65	4.25
1000	1	10	960	--	--	.52	.46	.44	1.13	4.22
	2	10	960	--	--	.62	.41	.49	1.07	4.59
	3	10	996	--	--	.70	.40	.40	1.07	3.19

Table 4.3.8

Summary of Standard Error Estimates for Various Sample Sizes and Ability Levels with a Homogeneous Item Pool (Test Length = 20 Items)

Sample Size	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
50	1	20	49	--	1.61	.45	.27	.44	.83	1.75
	2	20	49	--	.62	.56	.41	.33	.75	1.59
	3	20	50	--	2.53	.61	.17	.37	.63	1.52
200	1	20	194	--	1.83	.48	.34	.37	.70	1.60
	2	20	196	--	2.58	.47	.30	.39	.70	1.60
	3	20	198	--	2.64	.47	.30	.31	.69	2.03
1000	1	20	977	--	2.13	.46	.33	.33	.72	2.15
	2	20	984	--	2.09	.46	.34	.33	.68	2.01
	3	20	980	--	3.16	.53	.32	.33	.67	1.89

Table 4.3.9

Summary of Standard Error Estimates for Various Sample Sizes and Ability Levels with a Homogeneous Item Pool (Test Length = 80 Items)

Sample Size	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
50	1	80	50	1.50	.69	.30	.16	.18	.31	.59
	2	80	50	.75	.31	.21	.18	.20	.40	.79
	3	80	50	1.14	.56	.26	.15	.22	.34	.64
200	1	80	200	1.17	.46	.23	.18	.21	.36	.68
	2	80	200	1.00	.40	.21	.20	.22	.37	.69
	3	80	200	1.08	.47	.24	.17	.20	.35	.72
1000	1	80	1000	1.21	.49	.23	.19	.20	.34	.71
	2	80	1000	1.24	.49	.23	.19	.20	.35	.71
	3	80	1000	1.13	.44	.23	.20	.21	.33	.69

Table 4.3.10

Summary of Standard Error Estimates for Various Test Lengths and Ability Levels with a Homogeneous Item Pool (Sample Size = 50 Examinees)

Test Length	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
10	1	10	37	--	.66	.53	.51	1.43	1.05	2.71
	2	10	48	--	--	.79	.24	.48	1.59	--
	3	10	45	--	.80	.39	.42	.70	1.67	4.02
20	1	20	49	--	1.61	.45	.27	.44	.83	1.75
	2	20	49	--	.62	.56	.41	.33	.75	1.59
	3	20	50	--	2.53	.61	.17	.37	.63	1.52
80	1	80	50	1.50	.69	.30	.16	.18	.31	.59
	2	80	50	.75	.31	.21	.18	.20	.40	.79
	3	80	50	1.14	.56	.26	.16	.22	.34	.64

Table 4.3.11

Summary of Standard Error Estimates for Various Test Lengths and Ability Levels with a Homogeneous Item Pool (Sample Size = 200 Examinees)

Test Length	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
10	1	10	185	--	3.13	.41	.48	.41	1.16	4.03
	2	10	192	--	.52	.40	.65	.39	1.20	4.44
	3	10	179	--	3.89	.35	.60	.46	1.65	4.25
20	1	20	194	--	1.83	.48	.34	.37	.70	1.60
	2	20	196	--	2.58	.47	.30	.39	.70	1.60
	3	20	198	--	2.64	.47	.30	.31	.69	2.03
80	1	80	200	1.17	.46	.23	.18	.21	.36	.68
	2	80	200	1.00	.40	.21	.20	.22	.37	.69
	3	80	200	1.08	.47	.24	.17	.20	.35	.72

Table 4.3.12

Summary of Standard Error Estimates for Various Test Lengths and Ability Levels with a Homogeneous Item Pool (Sample Size = 1000 Examinees)

Test Length	Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
10	1	10	960	--	--	.52	.46	.44	1.13	4.22
	2	10	960	--	--	.62	.41	.49	1.07	4.59
	3	10	996	--	--	.70	.40	.40	1.07	3.19
20	1	20	977	--	2.13	.46	.33	.33	.72	2.15
	2	20	984	--	2.09	.46	.34	.33	.68	2.01
	3	20	980	--	3.16	.53	.32	.33	.67	1.89
80	1	80	1000	1.21	.49	.23	.19	.20	.34	.71
	2	80	1000	1.24	.49	.23	.19	.20	.35	.71
	3	80	1000	1.13	.44	.23	.20	.21	.33	.69

(a) Item Pool One—Effect of Sample Size. The results of the simulations for a fixed test length of 10 items, which are reported in Table 4.3.1, clearly show the lack of stability of the SEE Curves for all sample sizes. There was little improvement, if any, due to increasing sample size. This result, however, may be due to the limited amount of data considered since improvements were obtained in Item Pool Two and at other test lengths.

From examination of Table 4.3.2, which contains the results of the 20 item simulations, it is apparent that the SEE Curves were beginning to stabilize. Except at extreme values of the ability continuum, the results were nearly as good as those obtained with the larger sample size N=1000).

At a test length of 80 items, Table 4.3.2 clearly shows that SEE Curves are highly stable. Similar to the effect noted with test lengths of 20, the expected decrease in variation of the standard errors with increase in sample size, is apparent only at ability levels of -1, +1, and +2.

(b) Item Pool One—Effect of Test Length. Examination of the results reported in Table 4.3.4 indicate that, for samples of size 50, as test length increased, variation in the SEE Curves decreased at all ability levels.

Tables 4.3.5 and 4.3.6, which represent the results of the simulations for sample sizes of 200 and 1000, clearly show the following trends: (1) the most stable SEE Curves were obtained for the longest test length, and (2) for all ability levels, variation in the SEE Curves decreased as test length increased.

Table 4.3.13 presents a summary of the data found in Tables 4.3.1 to 4.3.6. Entries in this table are the standard deviations of the standard errors of estimate obtained across the three replications of the various studies. Standard deviations are reported for each test length-sample size combination across five ability levels. Also included in Table 4.3.13 is the average of the standard deviations across ability levels for each test length-sample size combination. It is this latter value that is the focus of the following discussion.

Several trends are apparent from examination of the average variation of standard errors: (1) the variation decreased as test length increased for all sample sizes, (2) when test length was fixed at 10 items, sample size had little or no effect on the stability of the SEE Curves, and (3) sample size, generally, had a noticeable effect on the stability of the SEE Curves.

Figure 4.3.1 contains three graphs illustrating the effect of test length and sample size on the stability of the SEE Curves at five ability levels. Each graph represents a plot of the values of the SEE Curves obtained when sample size was held constant and test length was varied. It is clear, from examination of these graphs, that sample size has little effect on the stability of SEE Curves of short tests (n=10). The effect of sample size on the stability of the standard errors was most apparent for the intermediate length test (n=20). For a long test (n=80) sample size showed the most pronounced effect when there was an increase from 50 to 200 examinees. An effect was also noticed when sample size was increased from 200 to 1000 examinees, however, the improvements in precision were more modest in size.

Table 4.3.13

Variation of Standard Errors of Estimates at Several Ability Levels for Different Test Lengths and Examinee Sample Sizes (Heterogeneous Item Pool)

Test Length	Sample Size	Ability Level[1] -2.0	-1.0	0.0	1.0	2.0	Average Variation Across Ability Levels
10	50	.68	.20	.39	.23	.50	.40
	200	.85	.10	.60	.50	.31	.47
	1000	.32	.04	.47	1.40	.19	.60
20	50	.36	.10	.01	.05	.11	.16
	200	.38	.06	.06	.04	.06	.12
	1000	.22	.05	.04	.02	.02	.09
80	50	.11	.04	.01	.05	.03	.06
	200	.04	.02	.02	.01	.00	.02
	1000	.01	.01	.00	.00	.00	.00

[1]Each entry in this section was obtained by calculating the standard deviation of standard errors of estimates across three replications for a particular test length and sample size.

(c) Item Pool Two—Effect of Sample Size. Table 4.3.7 presents the results of the simulations involving test lengths of 10 items. It should be noted that no values are reported for ability level -3 and also that the only complete set of values at ability level -2 are reported for a sample size of 200. Values obtained at these ability levels fluctuated greatly and so they are not reported (a similar explanation applies to other results not reported). In summary, there was a substantial improvement in the precision of SEE Curves for increasing sample sizes. In fact, the improvements in precision of SEE Curves due to sample size for test lengths of 20 and 80 items are also clear from a study of Table 4.3.8 and 4.3.9.

(d) Item Pool Two—Effect of Test Length. The results of this investigation are reported in Tables 4.3.10 to 4.3.12. These results are very similar to those obtained for item pool one and therefore will not be discussed to any great extent. It is important to note that for all sample sizes and at all ability levels there appeared to be a fairly consistent tendency for the stability of the SEE Curves to increase as test length was increased.

Table 4.3.14 summarizes the results reported in Tables 4.3.7 to 4.3.12. Data are arranged in Tables 4.3.14 in the same manner in which they were arranged in Table 4.3.13. Examination of the average variation across ability levels, indicates that for all test lengths, sample size has a noticable effect on the stability of the SEE Curves. In comparison to the results reported in Table 4.3.13, the effect of test length on the average variation across ability levels is not so apparent. The reason for this is the smaller variation observed for short tests with this particular item pool.

4.3.2 Effects of Statistical Characteristics of an Item Pool on Precision of SEE Curves

A comparison of the results reported in Tables 4.3.13 and 4.3.14, indicated that for tests of 20 and 80 items, the variation in the SEE Curves, averaged across ability levels, is very similar for both item pools. For test lengths of 10, the situation is quite different. In order to make the average variations across ability levels at this test length comparable for both item pools, these values were recomputed for item pool two, excluding the values obtained for ability level of -2. The recomputed average variation values are .33, .38, and .42 for sample sizes of 50, 200 and 1000 respectively. It is clear that, for short tests, the homogeneous item pool (pool one) resulted in smaller average variations than did the heterogeneous item pool. A second point worth noting, is that the heterogeneous item pool (pool two) provided more stable Standard Errors at an ability of -2 for test lengths of 10 or 20 items than did the homogeneous item pool. For test lengths of 80, the results appear to be about the same for both item pools. It should also be noted that the homogeneous item pool generally results in greater stability of Standard Errors for ability levels between +1 and -1 than did the heterogeneous item pool.

4.3.3 Relationship Between Test Length and SEE Curves in Two Typical Item Pools

Figure 4.3.2 contains two graphs, representing item pools one and two. These graphs show the relationship between test length and SEE Curves. Item parameters were used to derive the Curves

Table 4.3.14

Variation of Standard Errors of Estimates at Several Ability Levels for Different Test Lengths and Examinee Sample Sizes (Homogeneous Item Pool)

Test Length	Sample Size	Ability Level[1]					Average Variation Across Ability Level
		-2.0	-1.0	0.0	1.0	2.0	
10[2]	50		.17	.11	.41	.28	.24
	200		.03	.07	.03	.22	.09
	1000		.07	.03	.04	.03	.04
20	50	.78	.07	.10	.05	.08	.22
	200	.37	.00	.02	.04	.00	.09
	1000	.50	.03	.01	.00	.02	.11
80	50	.16	.04	.01	.02	.04	.05
	200	.03	.01	.01	.01	.01	.01
	1000	.02	.00	.00	.00	.01	.01

[1]Each entry in this section was obtained by calculating the standard deviation of standard errors of estimates across three replications for a particular test length and sample size.

[2]Standard deviations were not calculated for this test length at ability level -2 because of extreme fluctuations in the data.

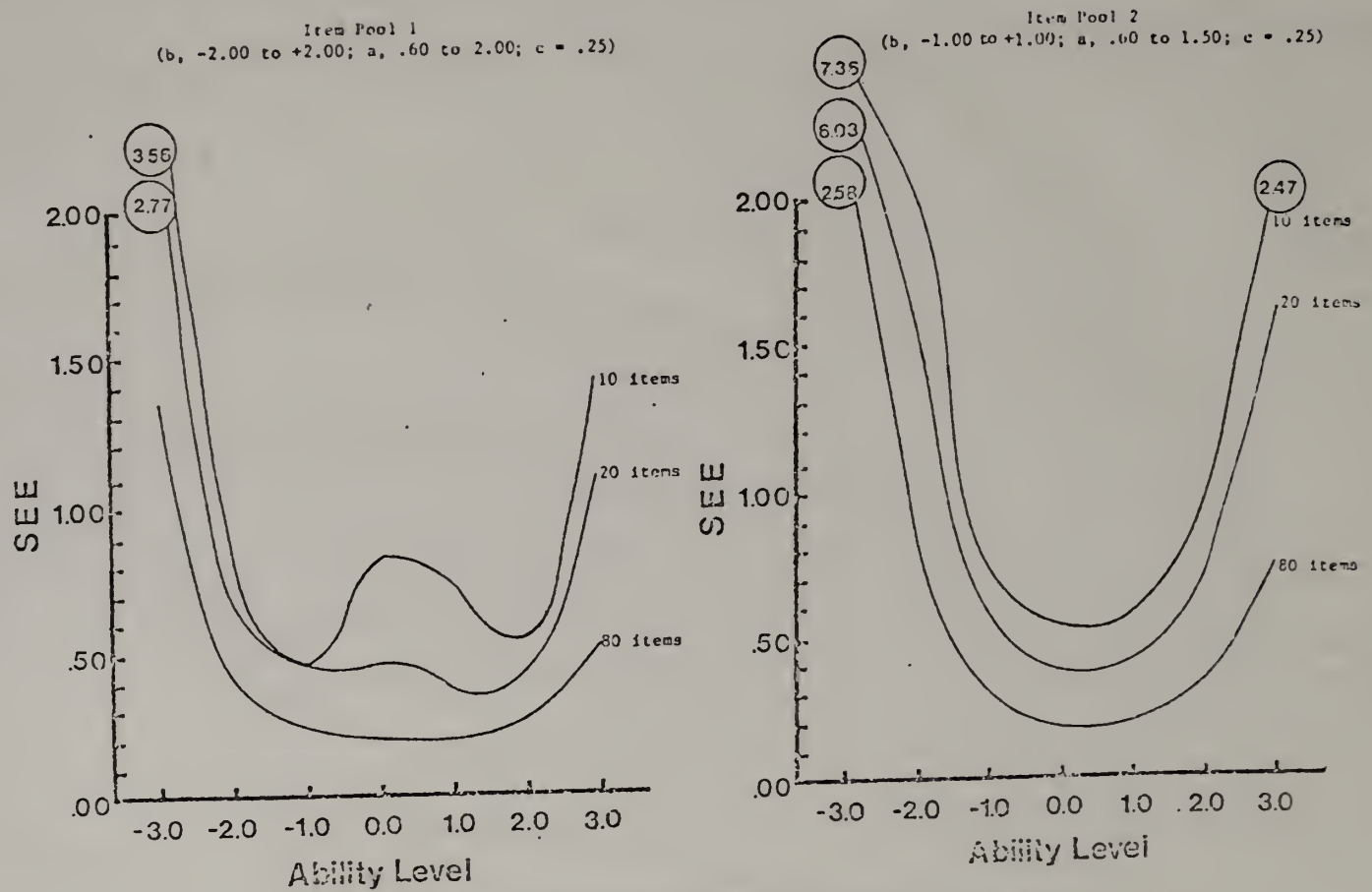


Figure 4.3.2 Standard errors of estimation associated with three test lengths at five ability levels and reported for two item pools.

rather than estimates of the item parameters. The trends in the results are generally what one would expect. The value of the figure is the information it provides to test developers who must determine a test length.

Test lengths of 10 and 20 items, drawn from the heterogeneous item pool (item pool one) do not show the expected U shaped pattern exhibited by the curves obtained for these test lengths when the simulation involved a homogeneous item pool. The slight distortion noted at the center of the ability distribution is due to the particular sample of items chosen. There are a few less items selected with difficulty values close to zero. It is quite apparent that the heterogeneous item pool provided smaller standard errors across a wider range of abilities than did the homogeneous item pool.

Further insight into the effect of the item pool on the size of the standard errors can be obtained by examination of the graphs presented in Figure 4.3.3. Each graph represents one of three different test lengths that was studied. The relationship between test length and SEE between +3 and -3 is graphed for both item pools on the same axes to facilitate comparison of the effect of the item pools. The decrease in the size of the standard errors as test length increases is quite evident for both pools. Also apparent is the fact that tests based on items drawn from the heterogeneous item pool provide greater precision over a wider ability range then do tests developed from the homogeneous item pool.

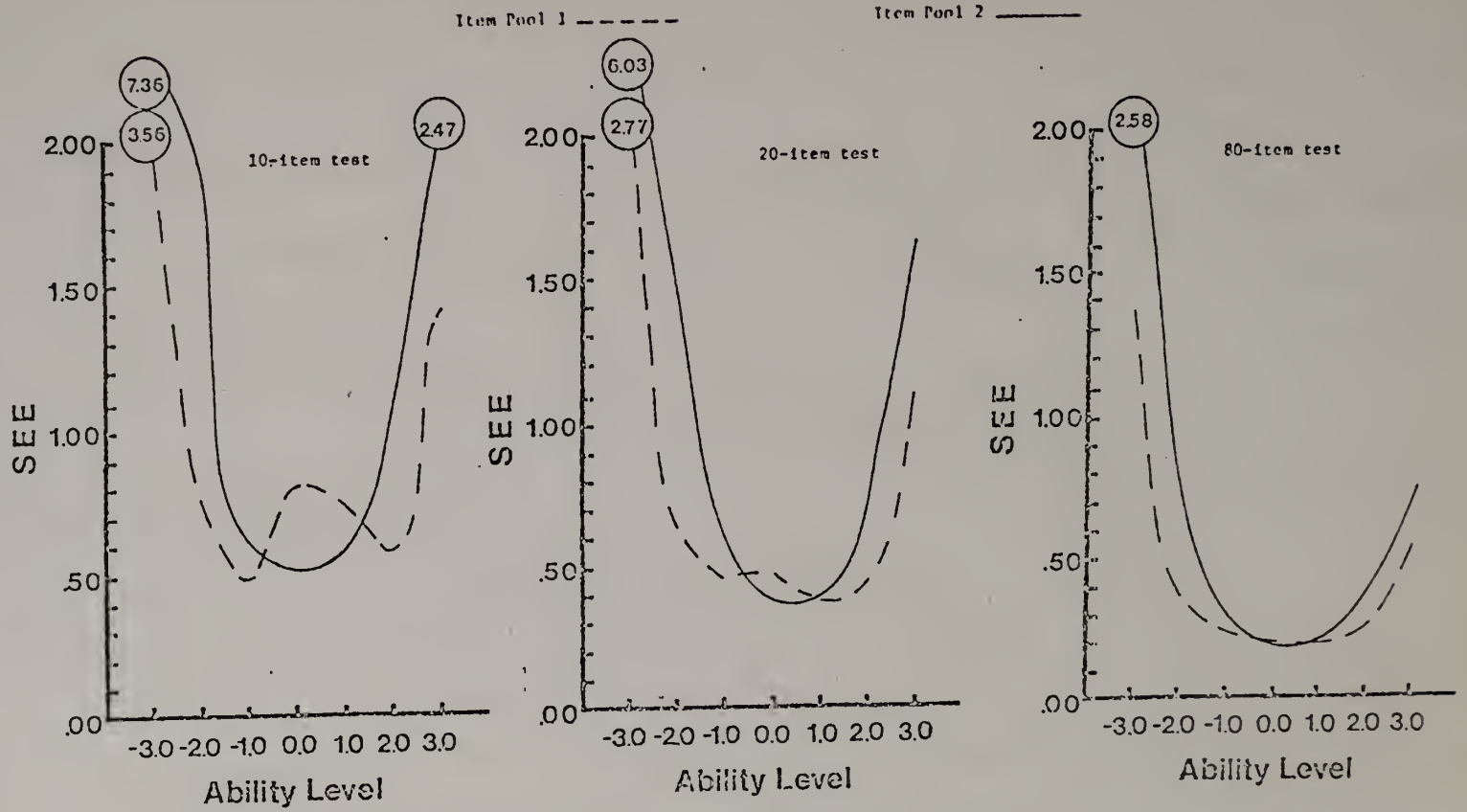


Figure 4.3.3 Standard errors of estimation associated with two item pools at five ability levels and reported for three test lengths.

4.4 Conclusions

A study along the general lines as this one is not going to reveal any major new results. It is well-known that the size of an examinee sample, the length of a test, and the characteristics of an item pool will have an important influence on the shape and stability of SEE Curves. The importance of this study is that it provides data concerning the size of improvements in SEE Curves relative to the three factors under investigation: (1) sample size, (2) test length, and (3) item pool characteristics. In this regard several conclusions seem warranted:

1. Both test length and sample size are extremely important factors in the precision of SEE Curves. (There were a small number of reversals in the results; no doubt this was due to sampling fluctuations.)

2. Precision of SEE Curves at the extremes of an ability continuum is very poor, even with large examinee sample sizes. The results are substantially better when tests are lengthened, even if the sample size is small (N=50).

3. The precision of SEE Curves would be acceptable in most instances if the Curves are based on 200 or more examinees with tests with at least 20 items. This recommendation holds if primary concern is with values of the Curves in middle regions of the ability continuum [-1 to +1].

4. Increases in examinee sample sizes from 50 to 200 produce sizeable improvements in the precision of SEE Curves. Gains in precision due to increasing a sample size from 200 to 1000 produce only modest gains in precision of the SEE Curves.

5. Similarly for test lengths, improvements in precision were substantially better when the change was from 10 to 20 items than 20 to 80 items.

The results of this study suggest that if an item pool is "typical," the stability of SEE Curves across readministrations of the test to similar groups of examinees will be quite good if the test includes at least 20 items, and if 200 or more examinees are used in deriving the item statistics.

CHAPTER V

A COMPARATIVE STUDY OF ITEM SELECTION METHODS UTILIZING LATENT TRAIT THEORETIC MODELS AND CONCEPTS

5.1 Introduction

Latent trait models offer a number of theoretical advantages to those interested in developing tests. Of course, the theoretical advantages of latent trait models will only be realized in practice if test data sets meet the restrictive assumptions imposed by the models being used (or considered for use) (Hambleton & Cook, 1977). Among these advantages are "sample-free" item statistics, "item-free" ability estimates, and a measure of the precision of ability estimation at different ability levels. This measure is called a "score information curve" (or a "test information curve" when the values of the item parameters are known). It is this last advantage and its potential usefulness in the test development process that is the focus of the research presented in this chapter.

Several excellent discussions of the use of information curves for the construction and evaluation of tests have been provided recently (Lord, 1977; Marco, 1977; Samejima, 1977; Wright, 1977). Still, the work conducted to date has not been addressed directly to the test practitioner who has an interest in using information curves to build tests.

The purposes of this particular investigation were as follows:

1. Using a typical item pool (where items are described by parameters in the three-parameter logistic test model), compare the score information curves for several item selection methods.

2. Compare the merits of several item selection methods for producing a scholarship exam and a test to optimally separate examinees into three ability categories.

The chapter has been divided into two parts to correspond to the purposes stated above. Part A, which focuses on a comparison of five item selection methods, is considered in section 5.2 to 5.4. Part B, which contains the investigation of item selection methods suited for two different testing purposes, is presented in sections 5.5 to 5.7. In addition, conclusions and practical implications of the work are discussed in a final section of the paper (5.8).

PART A

Comparison of Five Item Selection Methods

5.2 Purpose

The purpose of this part of the study was to investigate five possible item selection methods. In order to make the results of the five methods comparable, a fixed test length was used. Each method was used to select 30 items and the amount of information provided by the 30 selected test items, at five ability levels, -2, -1, 0, +1, +2, was calculated. The information curve obtained from each item selection method was then used to compare the methods. The five methods investigated were designated: (1) random, (2) standard, (3) middle difficulty, (4) up and down, and (5) maximum information. These procedures will be described in a later section.

An implicit assumption underlying these five methods was that the purpose of the test was to provide maximum information for ability levels ranging between -1 to +1, which is usually the case in practical test development situations.

5.3 Method of Investigation

5.3.1 Generation of the Item Pool

A computer program, DATAGEN (Hambleton & Rovinelli, 1973) was used to generate a "pool" of 200 test items. Each test item is described by the item parameters in the three-parameter logistic test model (item difficulty, item discrimination, and item pseudo-chance level). The item statistics are reported in Table 5.3.1. The average value and range of the item statistics were chosen to correspond to values which have been observed in practice (see, for example, Lord, 1968; Ross, 1966). Ability scores from a normal distribution (mean=0, sd=1) for 200 examinees were generated, and using the latent trait item statistics reported in Table 5.3.1, it was possible to simulate the item performance of the 200 examinees assuming the validity of the three-parameter logistic test model. With the availability of examinee item scores and total test scores, conventional item statistics (proportion-correct, and item-test score correlations) were calculated. These item statistics are also reported in Table 5.3.1.

Table 5.3.1

Item Pool Parameters and Item Information at Five Ability Levels
(b, -2.00 to +2.00; a, .19 to 2.00; c, .00 to .25)

	Item Parameters			Ability Level					Classical Statistics	
Item	b	a	c	-2	-1	0	1	2	p	r
1	.49	.49	.07	.04	.09	.14	.15	.11	.44	.36
2	-1.68	1.04	.25	.38	.39	.11	.02	.00	.93	.30
3	.09	1.11	.22	.00	.10	.55	.35	.07	.61	.48
4	1.73	1.70	.22	.00	.00	.00	.22	1.28	.30	.20
5	.81	1.44	.16	.00	.00	.25	1.10	.25	.37	.49
6	-1.41	1.32	.17	.42	.80	.15	.02	.00	.90	.41
7	1.38	.55	.13	.01	.03	.08	.15	.17	.39	.28
8	-.88	1.94	.19	.02	1.67	.43	.02	.00	.83	.62
9	1.45	.87	.12	.00	.01	.09	.35	.39	.27	.32
10	.47	1.21	.24	.00	.02	.39	.56	.13	.47	.48
11	.18	.32	.12	.03	.05	.06	.06	.05	.62	.22
12	.58	1.04	.25	.00	.03	.27	.46	.16	.50	.43
13	-.55	1.78	.22	.00	.62	.91	.06	.00	.78	.49
14	1.09	1.70	.23	.00	.00	.04	1.20	.40	.46	.37
15	1.01	1.39	.08	.00	.00	.22	1.19	.41	.28	.50
16	.88	.52	.12	.02	.06	.12	.15	.13	.45	.31
17	1.47	1.59	.04	.00	.00	.04	1.03	1.08	.14	.31
18	-.49	1.88	.04	.01	1.13	1.40	.08	.00	.71	.65
19	-1.00	1.45	.04	.28	1.39	.42	.04	.00	.82	.59
20	-1.80	.57	.18	.16	.14	.09	.04	.02	.89	.31
21	.73	1.21	.11	.00	.02	.37	.81	.24	.37	.50
22	.23	.72	.02	.06	.20	.35	.29	.13	.49	.44
23	.85	.96	.05	.00	.06	.34	.60	.29	.33	.45
24	-.37	1.10	.14	.03	.38	.63	.20	.03	.67	.58
25	1.21	.58	.17	.01	.03	.09	.17	.16	.38	.31
26	-.21	1.67	.19	.00	.20	1.36	.20	.01	.65	.57
27	-1.40	1.00	.04	.49	.60	.21	.05	.01	.86	.56
28	.82	.45	.09	.03	.06	.10	.12	.10	.53	.32
29	1.89	1.40	.13	.00	.00	.00	.23	1.11	.18	.30
30	-.11	1.70	.16	.00	.15	1.53	.26	.02	.63	.59
31	.27	1.94	.20	.00	.01	1.23	.64	.03	.49	.59
32	-.62	1.56	.22	.01	.67	.71	.07	.01	.77	.49
33	-.82	1.52	.09	.09	1.25	.58	.05	.00	.81	.55
34	1.93	.20	.09	.01	.02	.02	.02	.02	.48	.06
35	-1.54	1.34	.09	.72	.80	.13	.01	.00	.92	.36

Item	Item Parameters b	a	c	Ability Level -2	-1	0	1	2	Classical Statistics p	r
36	-1.63	.68	.14	.23	.24	.13	.05	.02	.85	.37
37	.08	1.36	.23	.00	.08	.79	.38	.05	.60	.59
38	-.46	1.39	.16	.02	.52	.84	.14	.01	.77	.49
39	1.17	.91	.03	.00	.04	.24	.55	.39	.25	.41
40	-1.29	1.92	.04	.58	2.04	.14	.01	.00	.87	.62
41	.18	.20	.12	.02	.02	.02	.02	.02	.53	.29
42	.34	1.58	.22	.00	.02	.75	.68	.06	.54	.54
43	-1.44	.36	.02	.09	.09	.08	.06	.04	.68	.33
44	-.49	1.40	.20	.01	.49	.78	.12	.01	.74	.56
45	-.28	.98	.01	.13	.48	.65	.26	.06	.64	.60
46	-1.90	.88	.11	.45	.32	.10	.03	.01	.91	.31
47	-.84	1.20	.15	.11	.72	.45	.08	.01	.78	.58
48	1.92	.71	.19	.00	.00	.03	.14	.25	.31	.31
49	1.62	.72	.20	.00	.01	.05	.18	.25	.32	.19
50	1.47	.64	.24	.00	.01	.06	.15	.18	.38	.31
51	1.77	1.40	.02	.00	.00	.03	.59	1.27	.13	.43
52	-1.26	1.56	.03	.59	1.49	.23	.02	.00	.87	.51
53	-1.13	1.62	.12	.22	1.50	.27	.02	.00	.89	.57
54	1.65	1.18	.23	.00	.00	.01	.28	.61	.33	.24
55	-1.15	1.59	.11	.25	1.44	.27	.02	.00	.87	.55
56	-.26	1.54	.01	.04	.71	1.52	.23	.02	.63	.69
57	1.52	.24	.15	.01	.02	.03	.03	.03	.48	.22
58	-1.88	1.22	.07	.91	.48	.08	.01	.00	.95	.30
59	-.33	1.26	.14	.02	.38	.81	.20	.03	.72	.57
60	.42	1.52	.22	.00	.01	.61	.74	.09	.55	.55
61	-.76	1.66	.12	.04	1.28	.65	.05	.00	.78	.59
62	-.44	.72	.21	.05	.19	.25	.14	.05	.68	.48
63	.11	.86	.15	.02	.14	.38	.29	.10	.55	.44
64	.90	1.82	.22	.00	.00	.09	1.57	.23	.44	.41
65	-1.21	1.14	.15	.27	.69	.25	.04	.01	.90	.43
66	.69	1.41	.13	.00	.01	.40	1.03	.20	.38	.50
67	.46	.42	.20	.02	.05	.08	.09	.07	.53	.29
68	.39	.40	.06	.05	.08	.10	.10	.08	.46	.33
69	-.07	1.83	.08	.00	.18	2.06	.29	.01	.63	.70
70	-1.66	.20	.02	.03	.03	.03	.02	.02	.68	.10
71	1.77	.20	.18	.01	.01	.02	.02	.02	.52	.21
72	.62	1.67	.16	.00	.00	.44	1.23	.13	.41	.46
73	-.08	.79	.23	.02	.14	.29	.21	.08	.65	.33
74	.79	1.47	.13	.00	.00	.30	1.19	.24	.32	.42
75	-.33	.69	.12	.07	.20	.26	.17	.07	.65	.34

Item	Item Parameters			Ability Level					Classical Statistics	
	b	a	c	-2	-1	0	1	2	p	r
76	.28	.61	.19	.02	.08	.17	.17	.10	.55	.42
77	1.25	.99	.05	.00	.02	.19	.61	.46	.27	.47
78	-.28	.70	.23	.04	.15	.23	.15	.06	.65	.45
79	1.90	1.55	.11	.00	.00	.00	.23	1.42	.22	.12
80	-1.27	1.84	.21	.21	1.47	.14	.01	.00	.88	.54
81	-.20	1.98	.22	.00	.13	1.79	.15	.01	.66	.58
82	-.41	.27	.22	.03	.03	.03	.03	.03	.59	.20
83	.57	1.28	.12	.00	.03	.52	.81	.17	.44	.46
84	.20	.55	.18	.03	.08	.14	.14	.09	.55	.36
85	-.81	.71	.19	.10	.24	.22	.11	.04	.75	.40
86	.78	1.53	.04	.00	.01	.50	1.47	.25	.31	.54
87	.62	1.73	.01	.00	.04	.98	1.59	.14	.33	.65
88	.11	1.51	.02	.00	.23	1.52	.54	.05	.51	.68
89	-1.24	1.21	.08	.41	.88	.26	.04	.01	.88	.46
90	-1.55	1.63	.06	1.02	1.05	.10	.01	.00	.90	.49
91	-.38	1.64	.07	.01	.67	1.37	.15	.01	.66	.65
92	.76	1.55	.15	.00	.00	.29	1.24	.21	.36	.51
93	.64	.67	.07	.02	.09	.23	.27	.17	.39	.35
94	-1.42	.90	.13	.32	.43	.18	.05	.01	.89	.31
95	-1.24	1.13	.03	.48	.83	.28	.05	.01	.86	.48
96	-.28	.54	.19	.05	.11	.15	.12	.07	.66	.27
97	-.94	.83	.18	.13	.34	.25	.09	.03	.82	.36
98	1.57	1.21	.25	.00	.00	.02	.32	.60	.35	.20
99	1.90	.72	.18	.00	.00	.03	.15	.26	.34	.22
100	-.58	1.06	.05	.12	.60	.58	.16	.03	.67	.54
101	-1.39	1.44	.11	.55	1.04	.17	.02	.00	.88	.51
102	1.00	1.43	.16	.00	.00	.13	1.07	.36	.38	.40
103	1.42	1.30	.23	.00	.00	.02	.49	.60	.32	.24
104	-.11	1.89	.17	.00	.11	1.84	.23	.01	.65	.55
105	.91	1.87	.10	.00	.00	.16	2.08	.27	.29	.51
106	1.18	1.50	.97	.00	.00	.11	1.29	.58	.27	.42
107	.53	1.17	.02	.01	.11	.69	.78	.19	.44	.50
108	-1.25	1.17	.20	.25	.67	.22	.04	.00	.85	.53
109	1.63	1.17	.14	.00	.00	.01	.46	1.11	.23	.29
110	1.22	.44	.06	.02	.05	.09	.12	.12	.28	.24
111	.75	1.46	.05	.00	.02	.52	1.30	.24	.33	.58
112	-1.90	.77	.12	.33	.26	.11	.03	.01	.89	.34
113	-.57	1.89	.02	.04	1.51	1.21	.07	.00	.71	.68
114	-.01	1.69	.00	.02	.42	2.07	.41	.03	.56	.65
115	1.10	.29	.09	.02	.03	.04	.05	.05	.45	.20

Item	Item Parameters			Ability Level					Classical Statistics	
	b	a	c	-2	-1	0	1	2	p	r
116	-.08	1.05	.18	.01	.18	.56	.28	.06	.69	.49
117	-.97	1.24	.25	.10	.67	.33	.05	.01	.85	.43
118	1.41	1.46	.01	.00	.00	.08	.99	1.08	.15	.45
119	-.24	.57	.24	.04	.10	.15	.12	.06	.64	.34
120	-.68	.42	.03	.10	.12	.12	.09	.06	.66	.30
121	1.12	.71	.00	.03	.09	.23	.36	.27	.24	.34
122	-.59	1.81	.06	.02	1.28	1.05	.06	.00	.74	.65
123	.51	.28	.10	.03	.04	.05	.05	.04	.50	.26
124	-1.04	.52	.19	.09	.13	.12	.07	.04	.80	.31
125	-.39	1.97	.20	.00	.35	1.43	.08	.00	.74	.57
126	.54	1.22	.18	.00	.03	.43	.66	.15	.44	.49
127	-.60	1.62	.23	.01	.66	.76	.07	.00	.78	.61
128	1.81	1.68	.18	.00	.00	.00	.20	1.42	.27	.25
129	-1.15	1.42	.03	.47	1.35	.31	.03	.00	.84	.58
130	-.91	.33	.04	.07	.07	.07	.06	.04	.66	.24
131	.42	1.07	.05	.01	.13	.62	.59	.16	.38	.55
132	1.41	.61	.25	.00	.01	.06	.14	.16	.46	.29
133	1.05	1.33	.24	.00	.00	.08	.75	.35	.40	.40
134	.18	.56	.19	.03	.08	.15	.14	.09	.66	.27
135	1.06	.39	.05	.03	.06	.09	.10	.09	.31	.30
136	-1.48	1.17	.09	.56	.71	.17	.03	.00	.86	.50
137	.44	.38	.15	.03	.05	.07	.08	.06	.52	.28
138	1.29	1.38	.08	.00	.00	.09	.98	.67	.22	.39
139	.15	1.45	.21	.00	.06	.89	.46	.05	.57	.53
140	-1.61	.48	.24	.09	.10	.08	.05	.02	.81	.31
141	-1.71	1.30	.25	.57	.50	.08	.01	.00	.92	.46
142	-.86	.80	.01	.25	.45	.33	.12	.04	.70	.47
143	1.62	.39	.01	.03	.05	.08	.11	.11	.30	.21
144	.16	.64	.20	.02	.09	.19	.18	.10	.58	.41
145	-.03	1.63	.01	.01	.42	1.90	.40	.03	.59	.65
146	-1.28	.28	.06	.05	.05	.05	.04	.03	.66	.11
147	-.29	1.08	.14	.03	.32	.62	.22	.04	.64	.53
148	1.71	.91	.18	.00	.00	.03	.24	.42	.31	.30
149	1.93	.86	.16	.00	.00	.03	.18	.39	.23	.12
150	-.87	.75	.18	.12	.27	.23	.10	.03	.76	.37
151	-.14	1.06	.25	.01	.16	.50	.25	.05	.63	.50
152	-1.75	1.96	.04	2.04	.75	.03	.00	.00	.96	.34
153	.14	1.18	.23	.00	.08	.58	.38	.07	.60	.48
154	-1.79	.87	.19	.35	.30	.11	.03	.01	.92	.30
155	-.88	1.87	.08	.07	2.03	.50	.02	.00	.86	.57

Item	Item Parameters b	a	c	Ability Level -2	-1	0	1	2	Classical Statistics p	r
156	.64	1.94	.24	.00	.00	.27	1.40	.09	.50	.45
157	-1.90	.65	.00	.30	.24	.12	.05	.02	.84	.38
158	-.74	.51	.19	.07	.12	.13	.09	.05	.68	.25
159	-1.73	.28	.18	.04	.04	.04	.03	.02	.76	.12
160	1.18	.79	.12	.00	.02	.14	.34	.29	.32	.38
161	-1.67	.63	.04	.25	.24	.13	.06	.02	.84	.33
162	.93	.38	.07	.03	.05	.08	.09	.08	.43	.29
163	.64	1.67	.06	.00	.01	.66	1.44	.15	.38	.60
164	1.45	.50	.17	.01	.02	.07	.11	.13	.44	.23
165	.90	.59	.11	.02	.06	.15	.20	.16	.41	.27
166	-.82	1.26	.00	.30	1.09	.57	.09	.01	.72	.58
167	-1.06	1.00	.17	.18	.52	.28	.07	.01	.81	.50
168	-.17	1.91	.12	.00	.20	2.02	.20	.01	.61	.62
169	.29	1.70	.12	.00	.03	1.19	.73	.05	.53	.60
170	.19	1.43	.05	.00	.15	1.25	.60	.07	.46	.65
171	.53	.25	.04	.03	.04	.04	.04	.04	.54	.27
172	-.11	1.28	.05	.02	.37	1.06	.34	.04	.55	.56
173	-1.04	.63	.04	.19	.27	.20	.10	.04	.76	.38
174	-1.19	.42	.01	.12	.13	.11	.07	.04	.69	.28
175	-1.67	1.57	.15	.92	.73	.07	.00	.00	.94	.32
176	1.86	.26	.23	.01	.01	.02	.03	.03	.44	.22
177	.00	1.96	.05	.00	.15	2.51	.35	.01	.59	.66
178	-.53	1.15	.23	.03	.38	.53	.13	.02	.73	.54
179	1.25	1.32	.14	.00	.00	.07	.80	.57	.36	.31
180	-.69	1.85	.16	.01	1.16	.74	.04	.00	.76	.56
181	.33	.24	.18	.02	.02	.03	.03	.03	.57	.17
182	-1.42	.86	.01	.44	.48	.21	.06	.01	.82	.38
183	-.40	.96	.00	.16	.52	.59	.22	.05	.66	.51
184	.40	.84	.03	.03	.17	.44	.41	.16	.43	.55
185	-1.92	1.23	.22	.67	.36	.06	.01	.00	.95	.35
186	.65	1.71	.07	.00	.01	.63	1.50	.15	.38	.57
187	-1.65	1.22	.09	.73	.63	.12	.02	.00	.90	.45
188	-.87	1.43	.24	.05	.82	.42	.05	.00	.82	.50
189	1.50	1.19	.16	.00	.00	.03	.46	.64	.26	.31
190	-1.86	.86	.25	.30	.26	.09	.02	.01	.88	.37

Item	Item Parameters b	a	c	Ability Level -2	-1	0	1	2	Classical Statistics p	r
191	-1.88	1.49	.24	.91	.42	.04	.00	.00	.96	.29
192	.10	.23	.07	.03	.03	.03	.03	.03	.51	.17
193	1.10	.63	.13	.01	.04	.12	.21	.19	.34	.29
194	.55	1.89	.15	.00	.00	.58	1.31	.08	.42	.48
195	-.96	1.17	.11	.18	.78	.39	.07	.01	.83	.48
196	-.16	1.25	.17	.01	.22	.80	.27	.04	.63	.52
197	-.11	1.60	.00	.04	.55	1.82	.33	.02	.56	.64
198	-1.02	.95	.08	.24	.55	.32	.08	.02	.82	.32
199	.88	.95	.11	.00	.03	.25	.52	.27	.40	.40
200	1.44	1.56	.20	.00	.00	.01	.61	.81	.34	.24

5.3.2 Item Selection Methods

(1) Random

Although it is unlikely that a practitioner with any degree of sophistication in the area of test development would select items at random, the results of such a process provide a base line for comparing the results obtained from other methods. To apply this method, a table of random numbers was used to select 30 test items from the pool.

(2) Standard

This method employed classical item statistics (item difficulty and item discrimination). Items were chosen such that their difficulties varied between .30 and .70. Of the total number of items with difficulty values falling in this range, the thirty items with the highest item discrimination parameters were chosen. The selected items had discrimination parameters that ranged between .53 and .70.

(3) Middle Difficulty

The 30 test items that provided the maximum amount of information at an ability level of 0.0 were selected from the pool.

(4) Up and Down

This method consisted of a three step process that was repeated until thirty items were selected. The three steps involved were to first, select the item from the pool that provided the maximum amount of information at an ability level of -1.0, next, proceed to an ability level of 0.0 and select the item that provided the

maximum amount of information at this ability level. The third step was to select the item at an ability level of +1.0 that provided the maximum amount of information. This three step process was repeated until 30 items were selected.

(5) Maximum Information

The fifth item selection method employed involved the averaging of information provided by each of the 200 items across three ability levels, -1.0, 0.0, and 1.0. The 30 test items providing the highest average levels of information across the three ability levels were selected for the test.

5.4 Results

The score information at five ability levels of interest for each of the five item selection methods are presented in Table 5.4.1. Table 5.4.1 also reports the number of the test items selected by each of the methods. Figure 5.4.1 provides a graphical representation of the score information curves resulting from the five item selection methods.

As was expected, the method employing a random selection of items provided less information than any of the other methods, at the ability levels of primary interest (-1.0, +1.0). It is interesting to note, however, that the score information curve resulting from this process is unimodal with maximum information provided at the center of the ability distribution. This result is a reflection on the nature of the item pool.

Table 5.4.1

Test Composition and Information Using Five Item Selection Methods

Item Selection Method	Selected Test Items (n=30)										Test Information				
											Ability Level				
											-2.0	-1.0	0.0	1.0	2.0
	1	2	9	11	13	15	45	54	56	58					
1. Random	65	71	76	81	82	93	97	108	118	121	2.61	5.99	12.43	10.14	4.57
	131	139	143	148	161	163	170	172	176	186					
	24	26	30	31	37	42	45	56	60	69					
2. Standard	81	86	87	88	91	100	104	111	114	131	.48	6.50	35.12	16.59	2.01
	147	163	168	169	170	172	177	184	186	197					
	13	18	26	30	31	37	38	42	44	56					
3. Middle	59	69	81	87	88	91	104	113	114	122	.27	11.38	40.68	10.24	.82
Difficulty	125	139	145	168	169	170	172	177	196	197					
	8	19	30	40	52	53	55	56	64	69					
4. Up and Down	80	81	86	87	104	105	106	111	113	114	2.84	19.00	27.48	18.06	2.35
	129	145	155	156	163	168	177	186	194	197					
	8	18	33	40	56	61	69	81	86	87					
5. Maximum	88	91	104	105	111	113	114	122	125	145	1.00	17.74	35.02	15.78	1.61
Information	155	163	168	169	170	177	180	186	194	197					

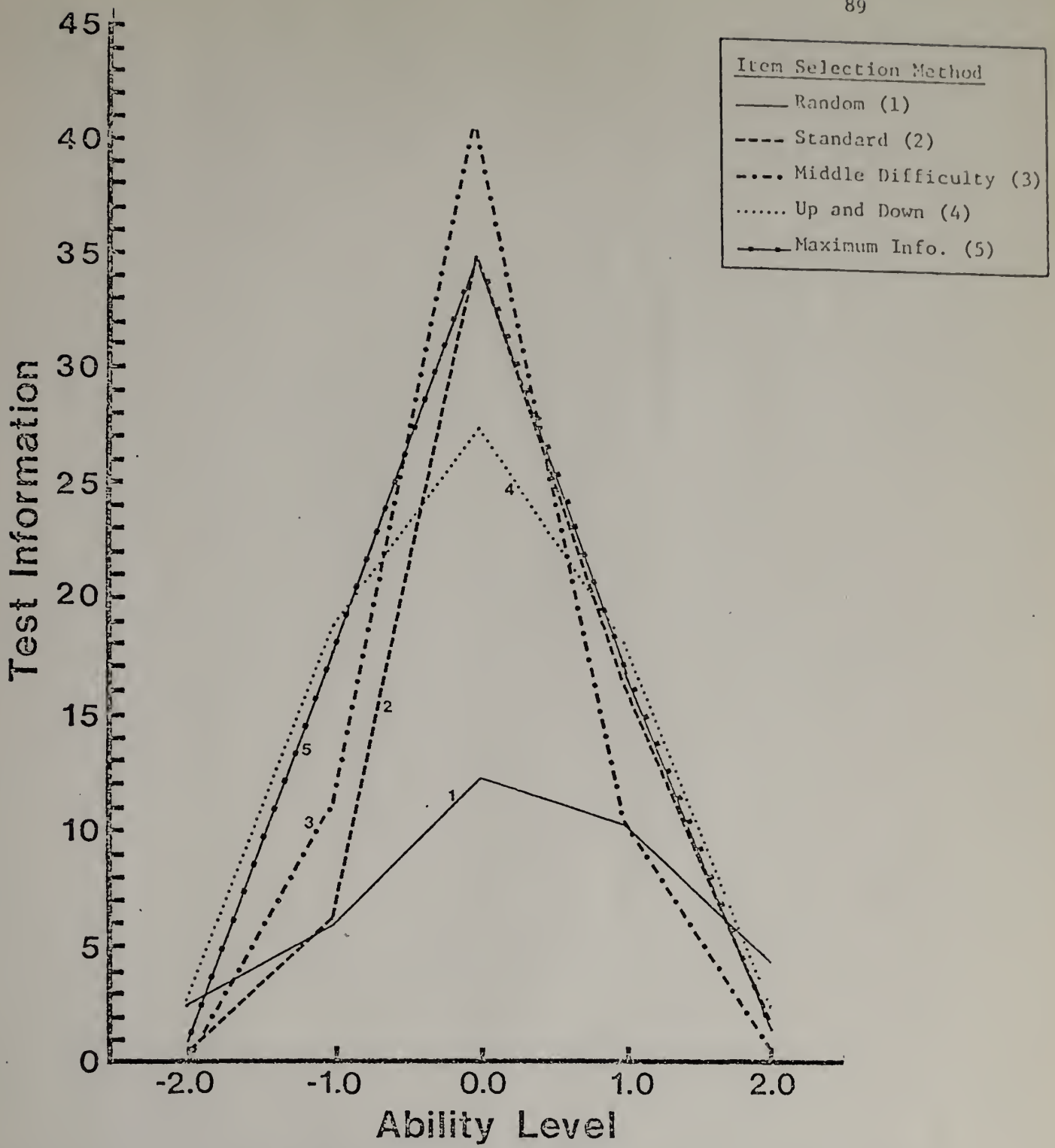


Figure 5.4.1. Test Information Curves Produced With Five Item Selection Methods [30 Test Items]

The "standard method" also resulted in a score information curve that provided maximum information for abilities at the center of the ability distribution. The amount of information provided at this point is considerably higher than that provided by the random approach. The information provided at an ability level of +1 is also considerably greater than that provided by the random selection method. This is not the case, however, for the amount of information provided at an ability level of -1.0. There is really very little difference between the two methods, in the values obtained at this ability level.

The third method, which involves selecting only those items that provided maximum information at an ability level of 0.0 resulted, as to be expected, in a score information curve that provides more information at this level than any of the other methods. This method also resulted in an appreciable amount of information at the two adjacent ability levels.

The "up and down" method, with the exception of the "random method," provided the least amount of information at θ=0.0 but it provides considerably more information at ability levels of -1.0 and +1.0 than did any of the other methods.

The "maximum information" method provided an appreciable amount of information at three ability levels (θ=-1.0, 0.0, 1.0). It did not provide as large an amount of information at an ability level of zero as did the "Middle Difficulty" method, however, it provided more information at the adjacent ability levels of +1.0 and -1.0 than did any of the other methods with the exception of the "Up and Down" method.

An interesting point to consider is the amount of overlap (in terms of percentage of items) that might be expected to result from each of these methods. Table 5.4.2 lists the number of overlapping items along with the percentage of overlap that this number represents. The smallest amount of overlap observed is four items. This occurred between the "Random" method and the "Up and Down" method. A surprisingly large amount of overlap was found between the "Standard" and the "Middle Difficulty" methods and the "Up and Down" and "Maximum Information" methods. Both of these pairs of methods had an overlap of 19 items (63.3%). In general the "Random" method appears to overlap least with the other item selection methods and the "Maximum Information" method seems to overlap the most.

PART B

Selecting Test Items to "Fit" Target Curves

5.5 Purpose

The second part of the study was designed to investigate item selection methods that are suited for a particular testing purpose. Two different testing situations were considered. The first situation, which was designated Case I, refers to the development of an instrument that is to be used for awarding scholarships, i.e., the maximum amount of information is desired at the upper end of the ability continuum. The second situation chosen to study (Case II) refers to the development of an instrument that will be used to make decisions at two different points on an ability continuum. An

Table 5.4.2

Overlap of Test Items Selected Using the Five Item Selection Methods
(Number of Common Test Items/Percent of Common Test Items)

Item Selection Method	Item Selection Method 2	3	4	5
1. Random	8 (26.7%)	6 (20.0%)	4 (13.3%)	5 (16.7%)
2. Standard	--	19 (63.3%)	14 (46.7%)	17 (56.7%)
3. Middle Difficulty	--	--	12 (40.0%)	17 (56.7%)
4. Up and Down	--	--	--	19 (63.3%)
5. Maximum Information	--	--	--	--

example of this type of instrument would be one that is used to award "passing" as well as "honors" grades to students.

For each situation (Case I and Case II), several item selection methods were developed and compared.

5.6 Method of Investigation

5.6.1 Case I

The investigation of Case I, the development of a scholarship selection instrument, began by establishing a target information curve. This was accomplished by specifying the size of the SEE that was considered desirable at each of the five ability levels ranging from -2.0 to +2.0. Using the relationship between the SEE and test information that was previously discussed (see Chapter IV), the amount of information required at each ability level was determined. The resulting target information curve is summarizes in Table 5.6.1 and presented graphically in Figure 5.6.1.

Four item selection methods were compared. These methods were designated: (1) Random, (2) Standard, (3) High Difficulty, and (4) Up and Down. Methods 3 and 4 are based on the use of item information curves. The "High Difficulty" method was one that involved choosing items that provided maximum information at an ability level of +1.0 (the ability level of primary interest). The "Up and Down" method involved the following steps: (1) choose the item that provides maximum information at an ability level of +2.0; (2) proceed to the adjacent ability level (+1.0) and select the item that provides maximum information at this ability level, (3) continue to work

Table 5.6.1

Target and Score Information Curves for the Two Test Development Projects

Case	Target/Method	Ability Level -2.0	-1.0	0.0	1.0	2.0	Number of Test Items Selected
I	Target	2.70	2.70	4.00	35.00	6.25	--
	Random	2.73	6.15	12.79	10.73	4.88	32
	Standard	.13	.60	4.25	18.14	17.18	32
	High Difficulty	.01	.44	13.18	35.01	9.29	32
	Up and Down	2.65	3.83	16.22	34.94	15.28	38
II	Target	4.00	25.00	4.00	25.00	4.00	--
	Random	3.47	8.65	13.65	11.36	5.29	36
	Standard	3.27	18.84	16.23	21.19	4.83	36
	Low-High Difficulty	4.77	25.26	15.39	24.86	5.27	36

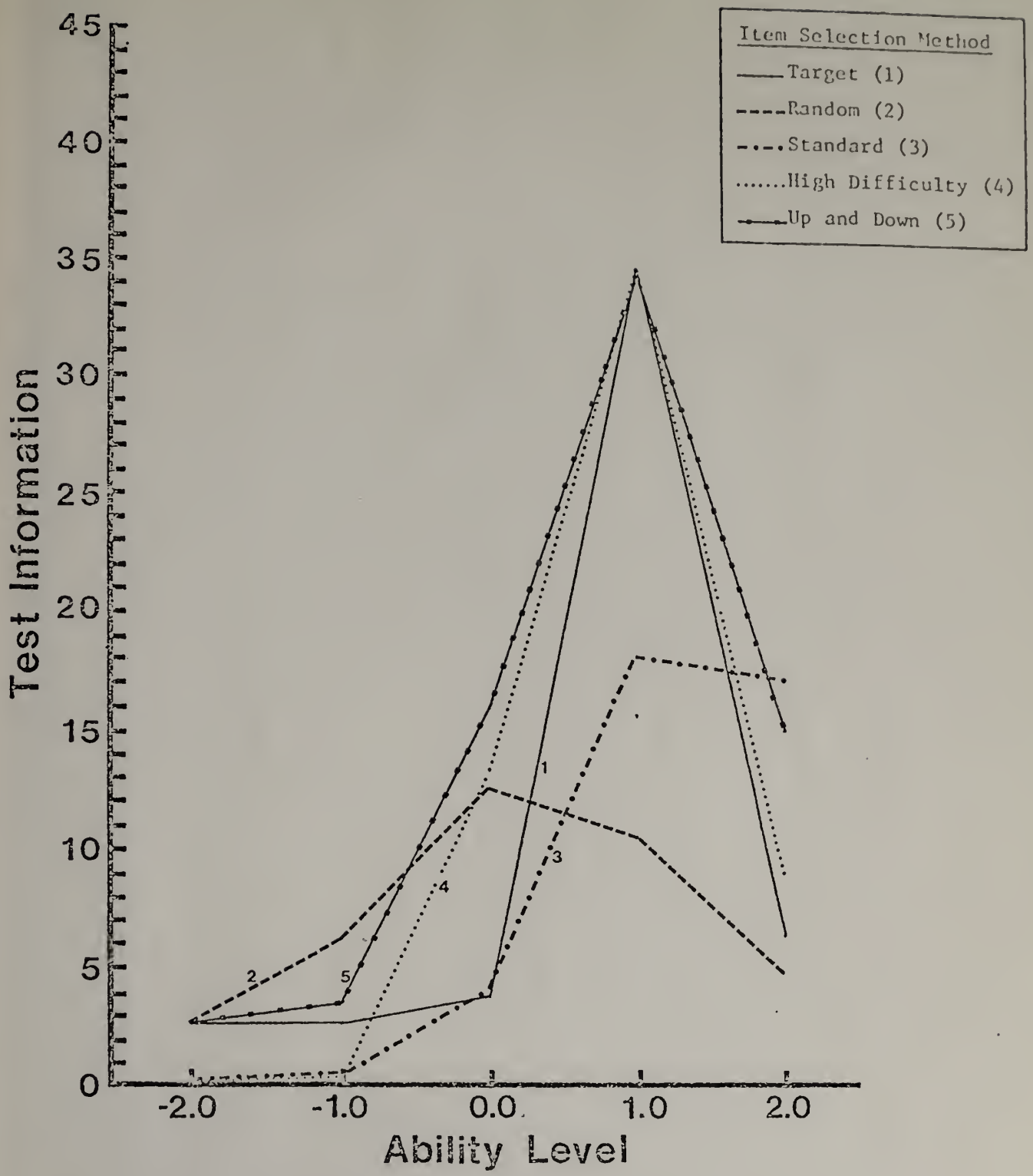


Figure 5.6.1. Scholarship Test Information Curves Produced With Five Item Selection Methods

down the ability continuum in this manner until an item is chosen that provides maximum information at an ability level of -2.0; (4) go back to an ability level of +2.0 and repeat the cycle. As the desired amount of information is obtained at a particular ability level, delete this ability level from consideration in the cycle.

The two remaining methods, which were not based on item information curves were similar to the random and standard methods described in an earlier part of this chapter, with the following exceptions: (1) The number of test items for each of these methods was set to be the same as the number of items required by the "best" of methods 3 and 4, (2) the specifications for the item difficulty values for the standard method were changed so that no item with an item difficulty value greater than .35 was chosen.

5.6.2 Case II

The target information curve for this testing situation was established by the same procedure described for Case I. The values for the resulting bimodal target information curve are summarized in Table 5.6.1 and presented graphically in Figure 5.6.2. It should be noted that maximum information is desired at two points on the ability continuum, -1.0 and 1.0.

Three item selections methods were compared for this testing situation. The only method based on the use of item information curves is the one designated "Low-High Difficulty." This method is similar to the "Up and Down" technique that was described previously and consists of selecting items alternately that provide maximum information at ability levels of +1.0 and -1.0. This back

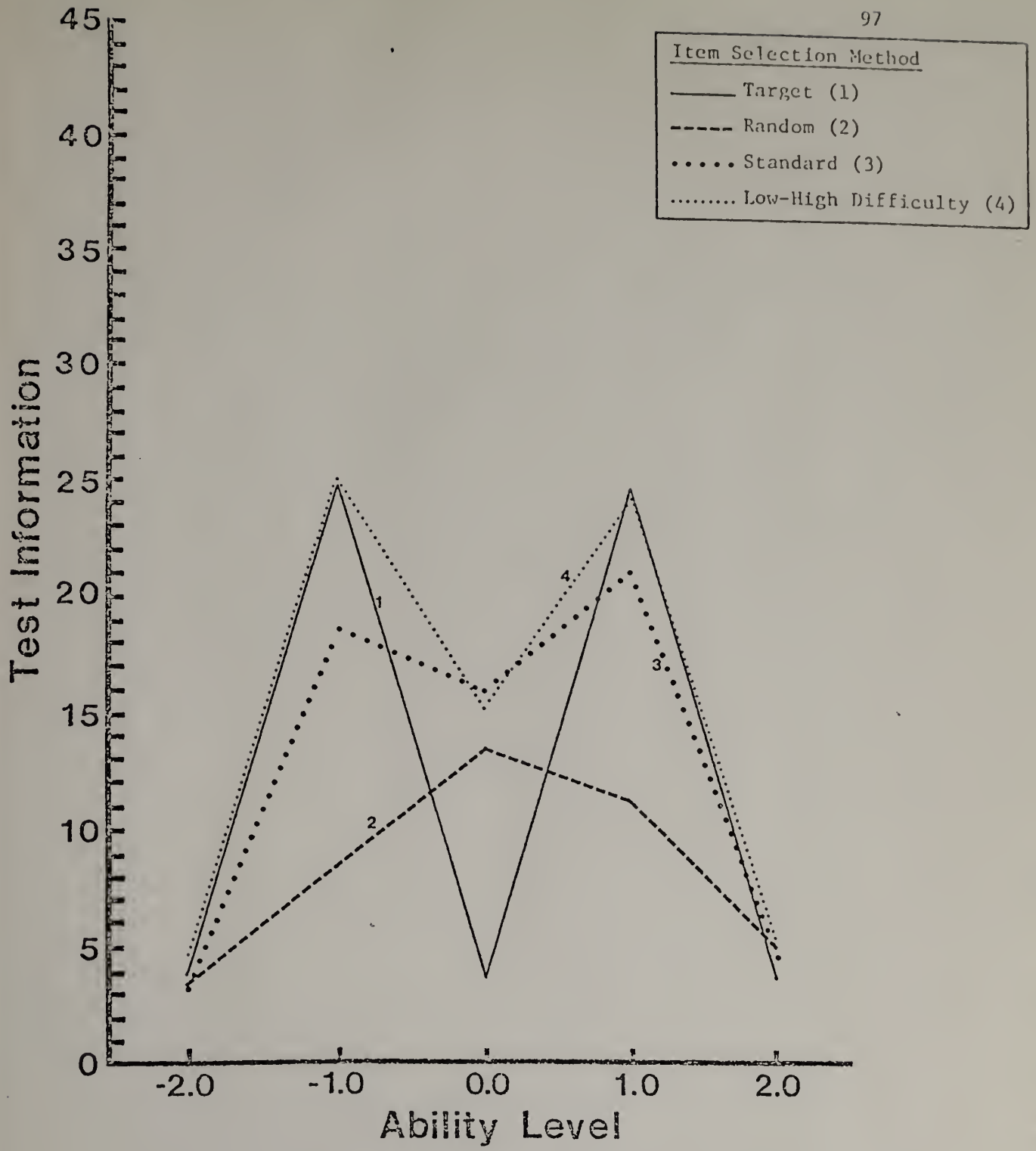


Figure 5.6.2. Bimodal Test Information Curves Produced With Four Item Selection Methods

and forth procedure is continued until the area under the target information curve is filled to a satisfactory degree. The random and standard methods are similar to those previously described. The number of items used for both of these methods was set to be the number of items required by the "Low-High Difficulty" method. The specifications for selecting items using classical item difficulty and item discrimination values were first to choose items with discrimination values greater than .40 and secondly, from this subset of items, to choose 18 items with difficulty values in the range of .70 to .90 and 18 items with difficulty values in the range of .20 to .40.

5.7 Results

5.7.1 Case I

The results of the four item selection methods are summarized in Table 5.6.1. A comparison of the two methods based on item information curves ("High Difficulty" and "Up and Down") shows that the "High Difficulty" method required six fewer items than the "Up and Down" method required to provide the desired amount of information at the ability level of interest (+1.0). The Random and Standard methods were clearly inferior. These results were certainly expected for the Random method but the dramatic difference between the amount of information at the ability level of interest obtained using the classical item statistics and that obtained using either of the other two methods is quite surprising. It is interesting to note that the "Up and Down" method provides maximum information over a broader

range of abilities than does the "High Difficulty" method, therefore it could possibly be a more appropriate technique for developing a selection instrument if moderate discrimination was also required at ability levels other than the one of major interest.

5.7.2 Case II

A summary of the results of the three item selection methods investigated is presented in Table 5.6.1. As expected, the "Random" method is totally inappropriate. The contrast between the "Standard" method and the method based on the use of item information curves is not as dramatic as in the Case I situation. Although clearly inferior, the results of the Standard method might possibly be useful in some situations. It is clear from Figure 5.6.2 that none of the methods provide score information curves that match the target information curve closely at points on the ability continuum other than those of major interest. However, the "Low-High Difficulty" method did provide a good test information-target information match at these points.

5.8 Conclusions

In all cases, the item selection methods based on either the random selection of items or the use of classical item statistics produced results inferior to those produced by methods utilizing latent trait model item parameters. And, the appropriateness of each method was situation specific. If maximum information is required at only one point on an ability continuum, it is clear that a method which chooses items that maximize information at this

particular point will be the best. If information is required over a wider range of abilities, methods involving averaging the information values across the ability levels of interest or choosing items in some systematic way that considers each point of interest on the ability continuum appear to be quite promising.

Only a limited number of methods and testing situations have been investigated, but the results indicate that it may be possible to pre-specify item selection methods that are situation specific and will enable a practitioner to develop a test quickly and efficiently without going through a lengthy trial and error process.

A variable not considered in this study was the effect of the item pool on the successful application of the methods investigated. It is quite possible that different results might have been found for item pools containing items with differing characteristics. Further research that consider other types of information based item selection methods as well as method-item pool interaction is certainly necessary before a complete set of generalizable guidelines can be developed.

CHAPTER VI

SUMMARY, CONCLUSIONS, AND IMPLICATIONS FOR FURTHER RESEARCH

This study had three purposes. The first was to systematically investigate the "goodness-of-fit" of the one-, two-, and three-parameter logistic models employing a practical criterion for assessment. Using computer-simulated test data, the effects of the following four variables were studied: (1) variation in item discrimination parameters, (2) the average value of the psuedo-chance level parameters, (3) test length, and (4) the shape of the ability distribution.

Simulated data was used so that it was possible to "know" examinee ability scores. These scores served as a criterion against which to judge the statistics derived from the three test models used to rank examinees. The rankings of examinees derived from each model (for each set of test data) were then compared to examinee "true" abilities. The Spearman rank difference formula was used to summarize the similarity between each pair of ranks (true abilities and estimates of ability from one of the models). Also reported are the average size of the discrepancies in the ranks for each group of 500 examinees.

The results of the study indicated: (1) the use of the item discrimination parameters as weights did not greatly improve the

proper ranking of examinees; (2) with short tests (n=20) the three-parameter model was more effective at ranking low ability examinees (this effect was found to be true for both normal and uniform ability distributions); (3) as test length increased, the three-parameter model continued to provide superior ranking of low ability examinees, however, the effect was not as great as with short tests; and (4) the number rights score proved to be about as effective for the ranking of high ability students as the more complicated scoring weights used in the two- and three-parameter logistic models.

The second study was designed to address two questions of importance and interest to the practitioner who wishes to use item and test information functions (or more specifically, a transformation of these functions that is referred to as the standard error of estimate of ability) as a means of selecting items for a test. The questions that the study was concerned with were:

1. What are the effects of examinee sample size and test length on the precision of standard error of ability estimation curves?

2. What effects do the statistical characteristics of an item pool have on the precision of standard error of ability estimation curves?

A computer simulation study was chosen as the mode of investigation for the two questions because of the large number of variables which were to be studied, and the need to "know" in some instances, the values of the item parameters.

Tests of three lengths were considered: 10, 20 and 80 items. Ability scores were simulated to be normally distributed (mean=0, sd=1). Three examinee sample sizes were studied: 50, 200, and 1000.

Ranges of parameter values for items in the two pools are shown below:

Item Parameter	Range of Values Pool One	Pool Two
Difficulty (b)	-2.00 to 2.00	-1.00 to 1.00
Discrimination (a)	.60 to 2.00	.60 to 1.50
Pseudo-Chance (c)	.25 to .25	.25 to .25

The results of the study can be summarized as follows:

Item Pool One

1. For a fixed test length of 10 items, the SEE Curves were clearly unstable for all examinee sample sizes.

2. When test length was increased to 20 items, the SEE curves began to stabilize, particularly at the center of the ability distribution.

3. Tests consisting of 80 items produced highly stable SEE Curves.

4. The expected increase in the stability of the SEE Curves with increase in examinee sample size, occurred only for test lengths of 20 and 80 items at ability levels of -1.0, +1.0 and +2.0.

5. For a long test (n=80) sample size had the most pronounced effect on the stability of the SEE Curves when there was an increase from 50 to 200 examinees.

6. For large samples (N=1000) test length had the most pronounced effect on the stability of the SEE Curves where there was an increase from 10 to 20 items.

Item Pool Two

1. Increasing sample size brought about a substantial improvement in the precision of SEE Curves.

2. For all ability levels and for all sample sizes, there was a consistent tendency for the stability of the SEE Curves to increase as test length increased.

Comparison of Item Pool One
With Item Pool Two

1. For short tests, item pool one resulted in more stable SEE Curves than did item pool two when variation across all ability levels was considered.

2. For test lengths of 10 or 20 items, item pool two produced more stable SEE values at an ability level of -2.0.

3. For long tests (n=80) the results were very similar for both pools.

The purpose of the third study was to investigate the following questions related to the development of item selection methodologies.

1. Using a typical item pool (where items are described by parameters in the three-parameter logistic test model), how does one develop alternate item selection methodologies and how do the score information curves that result from these methodologies compare?

2. Given a specific testing purpose such as producing a scholarship exam or a test to optimally separate examinees into three ability categories, how does one develop alternate item selection methodologies and how do the score information curves resulting from these methodologies compare?

The study was divided into two parts. Each part addressed one of the questions listed above. The first part of the study investigated five possible item selection methods. In order to make the results of the five methods comparable, a fixed test length was used. Each method was used to select 30 items and the amount of information provided by the 30 selected test items, at five ability levels, -2.0, -1.0, +1.0 , +2.0, was calculated. The information curve obtained from each selection method was then used to compare the methods. The five methods investigated were designated: (1) Random, (2) Standard, (3) Middle Difficulty, (4) Up and Down, and (5) Maximum Information.

The second part of the study was designed to investigate item selection methods that are suited for a particular testing purpose. Two different testing situations were considered. The first situation, which was designated Case I, refers to the development of an instrument that is to be used for awarding scholarships, i.e., the maximum amount of information is desired at the upper end of the ability continuum. The second situation chosen to study (Case II) refers to the development of an instrument that will be used to make decisions at two different points on an ability continuum. An example of this type of instrument would be one that is used to award "passing" as well as "honors" grades to students.

For each situation (Case 1 and Case II), several item selection methods were developed and compared. The Case I item selection methods were designated; (1) Random, (2) Standard, (3) High Difficulty, and (4) Up and Down. Only methods 3 and 4 are based on the use of item information curves. The Case II item selection methods were designated: (1) Random, (2) Standard, and (3) "Low-High Difficulty." Only the Low-High Difficulty method utilized item information curves.

The results of both part one and part two of this study indicated that item selection techniques based on either the random selection of items or the selection of items based on standard item statistics were inferior to the information function based item selection techniques.

A surprising amount of overlap in items chosen using the information based techniques was noted. Percentage of overlap varied

from 40% (Up and Down and Middle Difficulty) to 63.3% (Up and Down and Maximum Information).

The most effective methods found appeared to be those that focused on the selection of items suitable for specific ability levels. For the construction of a scholarship exam, the High Difficulty method proved to be the most effective. Whereas for the construction of an exam that discriminates well at two points on the ability continuum, the Low-High Difficulty method was the most efficient.

In conclusion, the results of the robustness study indicated that there are some sizable gains to be expected with modest length tests (n=20) in the correct ordering of examinees at the lower end of the ability continuum when three-parameter model estimates are used (as opposed to the number right score). The gains were cut roughly in half when the tests were doubled (n=40) in length. It was surprising that item discrimination parameters as weights had so little effect on the results. To the extent that the simulated data sets are typical of real data, it would appear that the application of latent trait models to the problem of "ranking" examinees is probably not worth the trouble except in those situations where gains of the size noted for lower ability examinees are important. The number right score does nearly as good a job of ranking examinees as the most complicated scoring methods.

The conclusions of the second study, which involved the investigation of the effects of test length, sample size and characteristics of the item pool on the stability of the SEE Curves were as follows:

1. Both test length and sample size are extremely important factors in the precision of SEE Curves. (There were a small number of reversals in the results; no doubt this was due to sampling fluctuations.)

2. Precision of SEE Curves at the extremes of an ability continuum is very poor, even with large examinee sample sizes. The results are substantially better when tests are lengthened, even if the sample size is small (N=50).

3. The precision of SEE Curves would be acceptable in most instances if the Curves are based on 200 or more examinees with tests with at least 20 items. This recommendation holds if primary concern is with values of the Curves in middle regions of the ability continuum [-1.0 to +1.0].

4. Increases in examinee sample sizes from 50 to 200 produce sizable improvements in the precision of SEE Curves. Gains in precision due to increasing a sample size from 200 to 1000 produce only modest gains in precision of the SEE Curves.

5. Similarly for test lengths, improvements in precision were substantially better when the change was from 10 to 20 items than 20 to 80 items.

The results of this study suggest that if an item pool is "typical," the stability of SEE Curves across readministrations of the test to similar groups of examinees will be quite good if the test includes at least 20 items, and if 200 or more examinees are used in deriving the item statistics.

The results of the study that addressed the development and comparison of item selection algorithms indicate that in all cases, the item selection methods based on either the random selection of items or the use of classical item statistics produced results inferior to those produced by methods utilizing latent trait model item parameters. It was also found that the appropriateness of each method was situation specific. If maximum information is required at only one point on an ability continuum, it is clear that a method which

chooses items that maximize information at this particular point will be the best. If information is required over a wider range of abilities, methods involving averaging the information values across the ability levels of interest or choosing items in some systematic way that considers each point of interest on the ability continuum appear to be quite promising.

There would appear to be at least three important implications for further research that are directly related to this study. The first focuses on the choice of criterion for assessing the "goodness-of-fit" of a latent trait model to a particular data set. The weaknesses of the commonly use X^2 test have already been explicated. Also mentioned is the fact that the criterion chosen for the goodness-of-fit study presented in this dissertation is appropriate only for norm-referenced testing situations. In the absence of appropriate criteria, practitioners are applying latent trait models, in many instances, to data that do not fit the assumptions of the models. The results of these mis-applications may have a serious effect on the observed test scores and ultimately the examinees who have been subjected to these tests. It is apparent that one of the most important contributions that could be made to the field of latent trait theory is the development of an adequate method for assessing goodness-of-fit.

The second area for further research is related to the study that addresses the stability of the SEE Curves. Because of the importance of these functions to the test development process, it is essential that data based on these functions be accurate. The

study presented in this dissertation was quite limited in terms of the number of variables that were investigated. The three major weaknesses of the study are related to the choice of computer programs employed, the limited variation between the two item pools that were chosen to be studied, and the ability distributions that were selected for study. Further research is certainly needed to compare the results obtained from such computer programs as LOGIST and BICAL with those obtained from Urry's program. It is also expected that further insight into the problem would be obtained if highly skewed ability distributions were employed and if item pools that differ from the characteristics of those studied were used.

The final area for further research concerns the study of item selection methodologies. If latent trait theory is to be implemented effectively these methodologies must be perfected and also simplified so that they may be applied in a routine manner. Investigations that involve a wide variety of testing purposes and different types of item pools are essential. The development of interactive computer programs based on appropriate methodologies would also greatly facilitate the application of latent trait theory to the test development process.

REFERENCES

Anderson, J., Kearney, G. E., & Everett, A. V. An evaluation of Rasch's structural model for test items. *British Journal of Mathematical and Statistical Psychology*, 1968, *21*, 231-238.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 1972, *37*, 29-51.

Cook, L. L., & Eignor, D. R. Considerations in the application of latent trait theory to objective based criterion-referenced tests. A paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Cook, L. L., & Hambleton, R. K. Application of latent trait models to the development of norm-referenced and criterion-referenced tests. *Laboratory of Psychometric and Evaluative Research Report No. 72*. Amherst, MA: School of Education, University of Massachusetts, 1978.

Dinero, T. E., & Haertel, E. Applicability of the Rasch model with varying item discriminations. *Applied Psychological Measurement*, 1977, *1*, 581-592.

Gulliksen, H. *Theory of mental tests*. New York: John Wiley and Sons, 1950.

Hambleton, R. K. An empirical investigation of the Rasch test theory model. Unpublished doctoral dissertation, University of Toronto, 1969.

Hambleton, R. K., & Cook, L. Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement*, 1977, *14*, 75-96.

Hambleton, R. K., & Rovinelli, R. A FORTRAN IV program for generating examinee response data from logistic test models. *Behavioral Science*, 1973, *18*, 74.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. Developments in latent trait theory: A review of models, technical issues, and applications. *Review of Educational Research*, 1979, in press.

Hambleton, R. K., & Traub, R. E. Information curves and efficiency of three logistic test models. *British Journal of Mathematical and Statistical Psychology*, 1971, *24*, 273-281.

Hambleton, R. K., & Traub, R. E. Analysis of empirical data using two logistic latent trait models. *British Journal of Mathematical and Statistical Psychology*, 1973, *26*, 195-211.

Hambleton, R. K., & Traub, R. E. The robustness of the Rasch test model. *Laboratory of Psychometric and Evaluative Research Report No. 42*. Amherst, MA: School of Education, University of Massachusetts, 1976.

Lawley, D. N. On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, 1943, *61*, 273-287.

Lawley, D. N. The factorial analysis of multiple item tests. *Proceedings of the Royal Society of Edinburgh*, 1944, *62-A*, 74-82.

Lazarsfeld, P. E. The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer et al., *Measurement and prediction*. Princeton: Princeton University Press, 1950.

Lord, F. M. A theory of test scores. *Psychometric Monograph*, 1952, No. 7.

Lord, F. M. An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, 1953, *18*, 57-75. (a)

Lord, F. M. The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 1953, *13*, 517-548. (b)

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 1968, *28*, 989-1020.

Lord, F. M. Estimating item characteristic curves without knowledge of their mathematical form. *Psychometrika*, 1970, *35*, 42-50.

Lord, F. M. Individualized testing and item characteristic curve theory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.) *Contemporary developments in mathematical psychology, Vol. II.* San Francisco: Freeman, 1974. (a)

Lord, F. M. Quick estimates of the relative efficiency of two tests as a function of ability level. *Journal of Educational Measurement*, 1974, *11*, 247-254. (b)

Lord, F. M. A survey of equating methods based on item characteristic curve theory. *Research Bulletin 75-13.* Princeton, NJ: Educational Testing Service, 1975. (a)

Lord, F. M. Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters. *Research Bulletin 75-33.* Princeton, NJ: Educational Testing Service, 1975. (b)

Lord, F. M. The "ability" scale in item characteristic curve theory. *Psychometrika*, 1975, *44*, 205-217. (c)

Lord, F. M. Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 1977, *14*, 117-138.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading MA: Addison-Wesley, 1968.

Lumsden, J. Test theory. *Annual Review of Psychology*, 1976, *27*, 251-280.

Marco, G. Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 1977, *14*, 139-160.

Panchapakesan, N. The simple logistic model and mental measurement. Unpublished doctoral dissertation, University of Chicago, 1969.

Pine, S. M. Applications of item response theory to the problem of test bias. In D. J. Weiss (Ed.), Applications of computerized adaptive testing. *Research Report 77-1.* Minneapolis, MN: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1977.

Rentz, R. R., & Bashaw, W. L. *Equating reading tests with the Rasch model, Volume I final report, Volume II technical reference tables.* Athens, GA: University of Georgia, Educational Research Laboratory, 1975.

Ross, J. An empirical study of a logistic mental test model. *Psychometrika*, 1966, *31*, 325-340.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, 1969, No. 17.

Samejima, F. A general model for free-response data. *Psychometric Monograph*, 1972, No. 18.

Samejima, F. A use of the information function in tailored testing. *Applied Psychological Measurement*, 1977, *1*, 233-247.

Tinsley, H., & Dawis, R. A comparison of the Rasch item probability with three common item characteristics as criteria for item selection. (Technical Report No. 3003), January 1972, Project No. NR 151-323, Personnel and Training Research Programs, Office of Naval Research. (ERIC Document Reproduction Service No. ED 068 516.)

Torgerson, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.

Urry, V. W. A Monte Carlo investigation of logistic test models. Unpublished doctoral dissertation, Purdue University, 1970.

Urry, V. Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement*, 1974, *34*, 253-269.

Weiss, D. J. The stratified adaptive computerized ability test. *Research Report 73-3*. Minneapolis, MN: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1973.

Weiss, D. J. Adaptive testing research at Minnesota: Overview, recent results, and future directions. In C. L. Clark (Ed.), *Proceedings of the First Conference on Computerized Adaptive Testing*. Washington, DC: United States Civil Service Commission, 1976.

Whitely, S., & Dawis, R. V. The nature of objectivity with the Rasch model. *Journal of Educational Measurement*, 1974, *11*, 163-178.

Wood, R. Response-contingent testing. *Review of Educational Research*, 1973, *43*, 529-544.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. *Research Memorandum 76-6*. Princeton, NJ: Educational Testing Service, 1976.

Wright, B. D. Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service, 1968.

Wright, B. D. Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 1977, *14*, 97-116.

Wright, B. D., Mead, R., & Draba, R. Detecting and correcting item bias with a logistic response model. *Research Memorandum No. 22*. Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1976.

Wright, B. D., & Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 1969, *29*, 23-48.